NOVA IMS Information Management School

DOCTORATE PROGRAM

Information Management

Specialization in Statistics and Econometrics

A Gaussian random field model for similarity-based smoothing in Bayesian disease mapping

Maria Helena Miranda Flores Baptista

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor in Information Management, under the supervision of Jorge M. Mendes

February, 2016

NOVA Information Management School
NOVA University of Lisbon

Prof. Doutor Jorge M. Mendes, Supervisor

A Gaussian random field model for similarity-based smoothing in Bayesian disease mapping

This document was created using the (pdf)LaTeX processor, based in the "unlthesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/unlthesis [2] http://www.di.fct.unl.pt

Acknowledgements

I am thankful to have the opportunity to express my deep gratitude to my advisor, Dr. Jorge M. Mendes, for his several years of support, encouragement and continuous challenge. This dissertation would not be possible without his guidance, and I consider myself fortunate to have had the opportunity to work with Dr. Jorge.

I would like to thank Dr. Ying MacNab for her help, advice and insight into the issues of this research and related work. Her knowledge is an invaluable resource to many critical points, and her extreme generosity of time and ideas are very much appreciated.

I would like to thank Dr. Duncan Lee for his helpful advice. His support was critical and I am thankful for his help in resolving important issues related to the use of the package CARBayes.

I would like to thank Dr. John Steward Huffstot for his help improving my English writing.

I would also like to thank Dr. Caldas-de-Almeida and Dr. Miguel Xavier for providing the World Mental Health Survey data as well as for their deep insight into the issues of mental health.

Personal thanks to follow student Ana Sofia Soares, who helped me through my years of coursework and made it enjoyable.

Finally, I would like to thank all my family for their continued support during the last years when I was more absent than I would have liked to be.

A dwarf standing on the shoulders of a giant may see farther than a giant himself
(Bernard of Chartres, 1115-1124).

I have been trying hard to see further by standing on the shoulders of Giants
(adapted from Newton (1676), who adapted it from Bernard of Chartres).

Abstract

Conditionally specified Gaussian Markov random field (GMRF) models with adjacency- or distance-based neighbourhood weight matrix, commonly known as neighbourhood-based GMRF models, have been the mainstream approach to spatial smoothing in Bayesian Disease mapping (DM).

In the present work, we propose a conditionally specified Gaussian random field (GRF) model with a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing in Bayesian DM. The model, named similarity-based GRF, is motivated for modeling DM data in situations where the underlying small area relative risks and the associated determinant factors do not varying systematically in space, and the similarity is defined by "similarity" with respect to the associated disease determinant factors.

The neighbourhood-based GMRF and the similarity-based GRF are compared and assessed via a simulation study and by two case studies, using new data on alcohol abuse in Portugal collected by the World Mental Health Survey Initiative (WMHSI) and the well-known lip cancer data in Scotland.

In the presence of disease data with no evidence of positive spatial correlation, the simulation study showed a consistent gain in efficiency from the similarity-based GRF, compared with the adjacency-based GMRF with the determinant risk factors as covariate. This new approach broadens the scope of the existing Conditional autocorrelation (CAR) models.

Keywords: Neigbourhood matrix, GMRF and GRF models, similarity-based smoothing, Besag-York-Mollié model (BYM) model, DM, Alcohol Abuse Disorder (AAD)

Resumo

Modelos especificados condicionalmente, denominados campos aleatórios Markovianos Gaussianos (CAMG), com matrices de vizinhanças ponderadoras baseadas em adjacências ou distâncias, comumente conhecidos como modelos CAMG baseados em vizinhanças, têm sido a aproximação principal utilizada no alisamento e inferência Bayesiana espacial em mapeamento de doenças.

Neste trabalho, propomos um modelo especificado condicionalmente, um campo aleatório Gaussiano com uma matrix ponderadora não espacial, mas baseada em similaridade para permitir o alisamento e inferência Bayesiana não espacial em mapeamento de doenças. O modelo, chamado CAG baseado em similaridade, foi motivado para a modelação em mapeamento de doenças em situações em que os riscos subjacentes em cada uma das pequenas áreas e seus factores determinantes associados não variam sistematicamente no espaço, e a similaridade é definida por "semelhança" com respeito aos factores determinantes associados à doença.

O modelo CAMG baseado em vizinhanças e o modelo CAG baseado em similaridade são comparados e avaliados através de um estudo de simulação e através de dois casos de estudo, usando os novos dados relativos ao abuso do álcool em Portugal recolhidos pelo estudo *World Mental Health Survey Initiative* e os dados, já bem conhecidos, do cancro dos lábios na Escócia.

Na presença de dados de doenças que não exibem uma correlação espacial positiva, o estudo de simulação mostra um constante ganho de eficiência pelo modelo CAG baseado em similaridade quando comparado com o modelo CAMG baseado em adjacência com os factores de risco como co-variáveis. Esta nova aproximação alarga o campo de utilização dos existentes modelos condicionais de autocorrelação.

Palavras-chave: Matriz de vizinhanças, CAMG e CAG, alisamento baseado em similaridade, modelo Besag-York-Mollié (BYM), Mapeamento de doenças, Abuso de Álcool.

Contents

List of Figures

List of Tables

Acronyms

AAD Alcohol Abuse Disorder.

BYM Besag-York-Mollié model.

CAPI Computer-assisted personal interview.

CAR Conditional autocorrelation.

CESOP Center for Public Opinion Studies and Polls.

CI Confidence intervals.

CIDI Composite International Diagnoses Interview.

DIC Deviance Information Criterion.

DM Disease mapping.

DSM - IV Diagnostic and Statistical Manual of Mental Disorders 4^{th} Edition.

EDF Estimated Degrees of Freedom.

ESEMeD European Study of the Epidemiology of Mental Disorders.

ESR Ecological-spatial regression.

GAM Generalized additive model.

GLM Generalized linear model.

GLMM Generalized linear mixed model.

GMRF Gaussian Markov random field.

GRF Gaussian random field.

HU Housing Unit.

ICAR Intrinsic conditional autocorrelation.

ICD-10 International Classification of Diseases 10^{th} Edition.

INE Statistics Portugal - Instituto Nacional de Estatística.

INLA Integrated nested Laplace approximations.

LCAR Localized Conditional Autoregressive model.

LLB Leroux-Lei-Breslow model.

MBYM Modified Besag-York-Mollié model.

McMC Markov chain Monte-Carlo.

NUTS3 Nomenclatura Comum das Unidades Territoriais Estatísticas.

OR Odds Ratio.

PAPI Paper-and-pencil assisted personal interview.

PSU Primary sampling unit.

RMSE Root-mean-square error.

SAE Small area estimation.

SMR Standardized morbidity ratio.

WHO World Health Organization.

WMH World Mental Health.

WMH-CIDI World Mental Health - Composite International Diagnoses Interview.

WMHSI World Mental Health Survey Initiative.

WT1 Weight on total sample.

WT2 Weight on reduced sample.

Chapter 1

Introduction

The availability of disease data in sets of non-overlapping and contiguous spatial areal units has increased over the last few decades. Concepts such as Small area estimation (SAE), DM and Ecological-spatial regression (ESR) are linked and are used in the context of the analysis of this type of data.

Firstly, we clarify the above concepts, and secondly, after focusing on DM, we apply several models to Portuguese Alcohol Abuse Disorder (AAD) data, collected by the WMHSI, as specified in Xavier et al. [94] and in Chapter 3.

The goal of DM is to estimate the spatial pattern in disease risk over a geographical region, so that small areas with elevated risk can be identified. Spatial DM models are being extensively used to describe geographical patterns of mortality and morbidity rates. Information provided by these models is considered invaluable by health researchers and policy-makers as it allows, for example, to allocate funds effectively in high risk areas, and/or to plan for localized prevention/intervention programmes.

The term DM was first used in Clayton and Kaldor [15]. It uses the spatial setting and assumes positive spatial correlation between observations, essentially "borrowing" more information from neighboring areas than from areas far away, smoothing local rates toward local neighboring values [84].

In cases of rare diseases and/or low populated areas, the classical estimators of the morbidity rates show high variability, and spatial DM models overcome that using the above mentioned characteristic. Models used in DM are usually Generalized linear mixed model (GLMM) formulated within a hierarchical Bayesian framework, and Poisson likelihood is often assumed for data in the form of counts of cases for each areal unit. Neighbourhood information is explicitly incorporated into the model by means of an appropriate prior specification. The seminal work of Besag, York and Mollié [9] provides a pair of area-specific random effects to model unstructured heterogeneity

(extra-Poisson variation) and spatial similarity. The BYM model is an extension of the Intrinsic conditional autocorrelation (ICAR) model, a well known GMRF prior in DM [9]. One important aspect of the CAR modelling is the definition of the so-called neighbourhood matrix, which characterizes the spatial structure of the data at hand, and is based on the concept of neighbours. Griffith [32] highlights the importance of the selected specification of the neighbourhood in spatial analysis of areal data.

The debate on the definition of neighbours can be traced back to Besag [8]. The author suggests that sites that comprise a finite system of closed irregular regions in the form of a mosaic, such as counties or states in a country, it will usually be natural to consider as neighbours of a given site, the sites that are adjacent to it. In a subsequent work, motivated by image analysis, where values from adjacent picture elements (pixels) influence the colour or grey-scale assigned to each pixel, Besag et al. [9] again define neighbours as those regions sharing a common boundary, the so-called adjacency-based GMRF matrix.

Best et al. [11] introduce a new definition of neighbourhood matrix based on distances between geographical centroids of local areas, the so-called distance-based GMRF matrix. Earnest et al. [24] propose examining the influence of different neighbourhood weight matrix structures on the amount of smoothing performed by the CAR model. By using four adjacency-based GMRF weight matrices and eight distance-based GMRF weight matrices, the authors report on considerable differences in the smoothing properties of the CAR model by the types of neighbourhood matrices specified.

Congdon [17] and Lee and Mitchell [50] work on cases in which one area is disparate from its neighbours. In these cases the global smoothing implemented by the CAR model may not be appropriate, and a local adaptive spatial smoothing is introduced. Contiguous areas showing clear discontinuities in the spatial patterns of health events are therefore considered conditionally independent. The authors move away from fixed adjacency-based GMRF matrices to estimated adjacency-based GMRF matrices.

Most of the research in DM is related with diseases resulting from environmental exposures, such as respiratory complications and cancer. Those extrinsic disease determinant factors are spatially smoothed, and using some kind of spatial proximity, either by adjacency or by distance, between areas in the definition of neighbours has therefore provided good results. In cases in which no spatial positive autocorrelation is displayed by the data, the neighbourhood matrix as it exists today may not be adequate. We propose a similarity-based GRF approach to replace the neighbourhood-based GMRF approach. The structure of the conditionals is maintained, but the smoothing and borrowing strength mechanisms are now based on the similarity of the areas, regardless of their relative location in space.

Our illustrating examples are two. Firstly, our motivating example is AAD occurrence in Portugal, a non-communicable disorder. Recent work [30] concludes that

alcohol abuse is a psychiatric disorder and not a socially defined consequence, and therefore its determinant factors are intrinsic. Secondly, we will use the much cited Scottish lip cancer data [11], which involves counts of lip cancer cases in the 56 districts of Scotland for 1975-1980. The choice of this dataset allows us to compare our results with those already published and assess adequacy of the proposed matrix when the disease determinant factors are extrinsic.

In Chapter 2, the DM definition highlighting the differences and common aspects among DM, SAE and ESR is provided. The main hierarchical Bayesian models, the CAR prior, and the actual neighbourhood matrices definitions are also presented. In Chapter 3 the data used for our motivating example are presented. This chapter ends with the presentation of the results achieved by the application of the several types of models (presented in Chapter 2) to the AAD data. In Chapter 4 the new approach for the weight matrix, the similarity-based GRF matrix, is explained. A simulation study is presented to demonstrate the performance of the proposed model. The two illustrating case studies are also presented. Finally, Chapter 5 presents a summary discussion.

Chapter 2

Disease Mapping

2.1 Introduction

The contents of this chapter is based on the paper: **Alcohol abuse disorder prevalence and its distribution across Portugal. A disease mapping approach** [7].

DM is linked to two other scientific areas: SAE and ESR. This chapter reviews similarities and differences among them. Bayesian hierarchical models are typically used in this context, using a combination of covariate data and a set of spatial random effects to represent the risk surface. The random effects are typically modeled by a CAR prior distribution, and a number of alternative specifications have been proposed in the literature. Four models will be assessed here.

Section 2.2 provides the DM definition highlighting the differences and common aspects among DM, SAE and ESR. Section 2.3 introduces the concept of Standardized morbidity ratio (SMR). Section 2.4 deals with the most common and widely used models for DM, providing some basic information on those, as well as some challenges and recent methodological advances. Section 2.5 presents the most widely used neighbourhood matrices in the DM context. Finally, Section 2.6 contains concluding remarks.

2.2 Small area estimation, Disease mapping and Ecological-spatial regression

DM joins together three different disciplines: statistics/biostatistics, epidemiology and geography. DM focuses on the challenge of obtaining reliable statistical estimates (statistics/biostatistics) of local disease risk based on counts of observed cases (epidemiology) within small administrative districts or regions (geography) coupled with

potentially relevant background information. DM goals are twofold: obtain statistically precise local estimates of disease risk for each region and maintain the regions "small" in order to keep the geographic resolution. The areas are not only small in size (relative to the area of the full spatial domain of interest), but are also small in terms of local sample size, resulting in deteriorated local statistical precision. To solve this problem the classical design-based solutions are often infeasible since the local sample sizes within each region, required for the desired level of statistical precision, are often unavailable or unattainable. The model-based approaches can help overcome this problem by the mechanism of "borrowing strength" across small areas to improve local estimates.

2.2.1 DM as a special case of SAE

Nowadays sample survey data are extensively used to provide reliable direct estimates of parameters of interest for the whole population. When it comes to getting the same estimates for domains of that population, and due to the small sample sizes in those domains, direct survey estimates are likely to yield unacceptably large standard errors. This makes it necessary to combine survey data collected from the small areas with auxiliary information from sources external to the survey. In this context, named as SAE, several indirect estimators have been extensively used. Some of the most common are the traditional indirect estimators based on implicit models, which include synthetic and composite estimators, and the Empirical Best Linear Unbiased Prediction approach. Most of these approaches also consider a contiguity matrix that describes the neighborhood structure between small areas "borrowing strength" from related areas to find more accurate estimates for a given area. The works of Rao [70] and Coelho and Pereira [16] provide respectively an overview of the foundations of SAE and a comparison between several traditional estimators and some proposed estimators using a Monte Carlo simulation.

DM is a special case of SAE, since the goal is to find reliable statistical estimates of local disease risk. As mentioned by Waller and Carlin [84] DM refers to a collection of methods extending SAE to directly utilize the spatial setting and assumed positive spatial correlation between observations. The data used are aggregated or averaged values at the small area level, representing disease incidence, prevalence or mortality rates, frequently not coming from surveys but coming from counts of disease cases from hospital admissions [52, 59]), counts of cancer cases or cancer deaths [9, 47, 82]), and mortality data [20, 59, 60]). In our motivating example we use counts of disease cases from a survey.

2.2.2 DM and ESR apply the same methodologies to reach different goals

By combining data from administrative registries and/or surveys with auxiliary data, DM goal is to predict area-level outcome summaries, to identify areas of elevated risk.

ESR uses the same type of data and the same methodologies but its objective is the estimation of associations between covariates and the disease cases.

Therefore, two common problems found in ESR are not of a concern in DM: (a) ecological bias and (b) the inclusion of spatially correlated errors changing the association between disease cases and fixed effects.

Ecological bias is the difference between estimated associations on ecological- and individual-level data [83]. Data used in DM and ESR, both for the number of cases and for the covariates are found rarely at individual-level, mainly due to confidentiality reasons, and therefore the association found at the aggregated level might not be the same if we would have used individual-level data. Aggregated data is usually designated as areal data [5]. The objective of DM is not to estimate the associations between the cases and the covariates or to improve predictions, and therefore ecological bias is not a concern (for more details on the subject see Wakefield and Lyons [83]).

The inclusion of spatially correlated errors, changing the association between disease cases and fixed effects, has been studied by Wakefield [82] and Hodges and Reich [36]. Often the study of ESR has provided estimates of the fixed-effect coefficients substantially different from those of ecological regressions. ESR is an ecological regression augmented with the inclusion of random effects modeled by a globally smooth conditional autoregressive model. If the covariates are also globally smooth, collinearity problems might change dramatically the coefficients of the fixed-effects. As before, the coefficients of association are not of direct interest in DM, and therefore this aspect is not a concern.

2.3 Standardized morbidity ratio

As said before, in DM we typically have count data per area. Those area count data are thought of as random variables. To assess the disease risk of each area the number of cases expected needs to be calculated. Those expected disease count data are thought of as fixed and known functions of the number of people in risk in each area. If we expect that the disease rate is constant in every area, those expected count data correspond to a kind of "null hypothesis". This process is called the age standardization process.

The age standardization process, as defined in Waller and Gotway [85], can be direct or indirect. The choice between direct and indirect standardization is usually defined by the type of data available. Is direct when age-specific rates for each of the small areas is available. Is indirect when those age-specific rates are only available for the whole population.

The age standardization process, can also be internal or external. Is internal when the survey used to collect the disease cases is also used to calculated the age-specific rates. External standardization only occurs when standard tables of age-specific rates for the disease are available. As mentioned in Banerjee et al. [5] internal standardization is "cheating" in some sense, since "a degree of freedom is lost" by estimating the

age-specific disease rate from the current data.

Accordingly, for an internal indirect standardization, the following notations and definitions are introduced:

a. Y_k the random variable representing the number of observed cases (y_k) in each k age group;

b. n_k representing the number of people at risk in each k age group;

c. $r_k = \frac{y_k}{n_k}$ representing the observed prevalence proportion for each k age group;

d. n_{ik} representing the number of people at risk in each k age group in the i^{th} small area;

e. E_{ik} and y_{ik} representing the expected and observed number of cases for the k age group in the i^{th} small area, respectively, where $E_{ik} = r_k n_{ik}$;

f. $E_i = \sum_k r_k n_{ik}$ and $y_i^* = \sum y_{ik}$ representing the total number of expected and observed cases in the i^{th} small area, respectively;

g. $SMR_i = \frac{Y_i}{E_i}$, the *standardized morbidity ratio*, representing the risk of each i^{th} small area. A value of SMR greater (less) than one indicates that the area i has a higher (lower) than average disease risk. If the $SMR_i = 1.15$, it can be said that area i has a 15% increased risk of the disease.

The use of SMRs can create various challenges, namely:

1. Because SMRs are given as ratios, is possible that to large changes of risk estimate only small changes are produced in the expected value of cases;

2. Zero SMRs do not distinguish variation in the corresponding expected count;

3. The standard errors of SMRs are inversely proportional to the expected number of cases.

In frequentist terms, the SMR is the maximum likelihood estimate of the disease risk in each small area, with a $\hat{VAR}(SMR_i) = Y_i/E_i^2$. This permits calculation of traditional Confidence intervals (CI) for the risk, as well as hypothesis tests. Assuming that $\log SMR_i$ is normally distributed, an approximate 95% CI for the risk is:

$$(SMR_i \exp(-1.96/\sqrt{Y_i}), (SMR_i \exp(1.96/\sqrt{Y_i})$$

2.4 Disease mapping models

DM methodologies are explained in Waller and Carlin [84] and Banerjee et al. [5]. DM methodologies for areal data are usually divided in frequentist methods and hierarchical Bayesian models [5]. To provide a wide comparison of methods, Lawson et al. [46] presents some preliminary results concerning the goodness-of-fit of a variety of DM methods applied to simulated disease incidence data. These simulated models cover simple risk gradients and more complex true risk structures, including spatial correlation. Authors conclude that full Bayesian hierarchical models are the most robust across a range of diverse models.

A number of hierarchical Bayesian models have been proposed in the literature, including the following two, which have been widely used: a) the model developed by Besag, York and Mollié [9]), from now on designated as BYM model and b) the model developed by Leroux, Lei and Breslow [55], from now on designated as LLB model. These two models will be used in Section 3.5. Best, Richardson and Thomson [12] review the main classes of Bayesian models, among which the BYM model is included (but not the LLB model) and conclude that the BYM model has good properties for modeling a single disease and "appears to be the only fully Bayesian spatial model to have been used in published applications of disease mapping outside of the statistical literature" (page 57). Recently, some authors [47, 59] published comparisons between hierarchical Bayesian models and both conclude that the LLB model is the best overall, because it produces consistently good results across a range of spatial correlation scenarios, is more parsimonious on parameters, and has less undesirable features (this subject will be further developed in Subsection 2.4.1).

One of the challenges posed at the DM level arises from its basic goal, the smoothing of local rates toward local neighboring values. When real discontinuities exist between neighboring areas, the models will lead to oversmoothing blurring the edges, which may not be appropriate. If the goal is to identify boundaries or regions of rapid change, the methods of *boundary analysis* or *wombling* need to be applied. For more detail see the recent works of Lee and Mitchell [49, 50].

A general formulation for the first level of the hierarchical Bayesian models used in DM is given by

$$Y_i|E_i, R_i \sim Poisson(E_i R_i) \text{ for } i = 1, ..., n,$$

$$\ln(R_i) = \mu + \mathbf{x_i^T}\boldsymbol{\beta} + \phi_i, \tag{2.1}$$

If E_i is not too large (as it is the case of rare diseases) or the regions i are sufficiently small, the usual model for the Y_i is a Poisson model [5]). In the model, R_i denotes the risk of disease in area i, which is modeled by an intercept term μ, a set of p covariates $\mathbf{x_i^T} = (x_{i1}, ..., x_{ip})$ multiplied by the corresponding vector of regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, and a random effect ϕ_i. The random effects are included to model any overdispersion and/or spatial correlation that might remain in the data

after having being accounted for by the included covariate information. Most studies of this type show overdispersion, meaning that $Var[Y_i] > \mathbb{E}[Y_i]$, which has several possible causes: subject heterogeneity; correlation between individual responses; omitted unobserved variables; and/or excess zero counts. Inference for this type of model is based on Markov chain Monte-Carlo (McMC) simulation, using a combination of Gibbs sampling and Metropolis-Hastings steps and more recently using Integrated nested Laplace approximations (INLA) [76].

The random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are usually modeled by the class of CAR prior distributions [5] , which are a type of GMRF [35]. Instead of a specification of a single multivariate distribution $f(\boldsymbol{\phi})$, the above models are specified by a set of univariate full conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. To determine the spatial correlation between the random effects, is usually used the neighborhood matrix $\mathbf{W}$, which is a binary $n \times n$ matrix, with elements w_{ji}:

$$w_{ji} = \begin{cases} 1, & \text{if } j \sim i \\ 0, & \text{otherwise,} \end{cases}$$

where $j \sim i$ represents contiguous areas, and therefore j and i are considered neighbors. Other *adjacency-based* weights are available but are much less widely applied [84]). If two areas are neighbors their random effects are correlated, while non-neighboring areas are modeled as being conditionally independent given the remaining elements of $\boldsymbol{\phi}$. We will return to this subject in Chapter 4.

2.4.1 BYM model

The BYM model combines the ICAR with an additional set of independent random effects.

The full conditional distributions of ICAR, as proposed by Besag et al. [9] are given by

$$u_i|\boldsymbol{u}_{-i}, \sigma^2 \sim N\left(\frac{1}{n_i}\sum_{j\sim i} u_j, \frac{\sigma^2}{n_i}\right). \tag{2.2}$$

The conditional expectation of u_i is equal to the mean of the random effects in neighborhood areas, while the conditional variance is inversely proportional to the number of neighbors n_i. The variance parameter σ^2 controls the amount of variation between the random effects. The ICAR model has three main drawbacks:

1. its simplicity turns it into a very restrictive prior. Its single parameter does not determine the strength of the spatial correlation (for example multiplying each u_i by 10, will only increase σ^2 leaving the spatial correlation unchanged). If data are weakly correlated, the ICAR is not the most appropriate model [47];

2. the joint distribution for $f(\mathbf{u})$ corresponding to (2.2) is improper (it does not determine a legitimate probability distribution, one that integrates to 1). Nevertheless, this is easily solved by enforcing a constrain such as, $\sum_{j=1}^{n} u_j = 0$, which

can be *numerically* imposed by recentering each sampled **u** vector around its own mean following each Gibbs iteration [5];

3. according to MacNab [59] the ICAR has an undesirable *global* (*i.e.* large-scale) property of tending to a negative pair-wise risk dependance as the "spatial proximity" of the two regions is further apart.

The BYM model defines ϕ in (2.1) by

$$\phi_i = \theta_i + \psi_i, \tag{2.3}$$

$$\theta_i|\sigma_\theta^2 \sim N(0, \sigma_\theta^2),$$

$$\boldsymbol{\psi} = (\psi_1, ..., \psi_n)|\mathbf{W}, \sigma_\psi^2 \sim ICAR(\mathbf{W}, \sigma_\psi^2),$$

where **W** is defined in Section 2.4. More details on the BYM model are provided by Lee [47] and Besag et al. [9].

The set of random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ is independent between areas. Different strengths of spatial correlation can be represented by varying the relative sizes of the two components $(\boldsymbol{\theta}, \boldsymbol{\psi})$. In practice, it will often be the case that either $\boldsymbol{\theta}$ or $\boldsymbol{\psi}$ dominates the other depending upon the strength of the spatial structure and the relative sizes of σ_θ^2 and the σ_ψ^2. This flexibility is also a disadvantage, as each data point is represented by two random effects while only their sum $(\theta_i + \psi_i)$ is identifiable. In order to attain model identification and achieve convergence when McMC is used, at least one considerably informative hyper prior has to be assumed either for σ_θ^2 or σ_ψ^2. Several authors have studied this aspect [5, 84]), and MacNab [59] implemented a model that can "attain model identifiability, allow the data to inform risk decomposition, and facilitate principled attribution of the relative risk variability to spatially varying *clustering* effects and randomly varying *heterogeneity* effects based on the *given data*" (page 66), hereafter called Modified Besag-York-Mollié model (MBYM). This model replaces (2.3) by

$$\boldsymbol{\phi} = \sqrt{\lambda}\boldsymbol{\psi} + \sqrt{1-\lambda}\boldsymbol{\theta},\ \boldsymbol{\psi} \perp \boldsymbol{\theta},\ \lambda \in (0,1). \tag{2.4}$$

One interpretation of the above is that it represents a re-parameterized BYM prior with $\sigma_\psi^2 = \lambda\sigma^2$ and $\sigma_\theta^2 = (1-\lambda)\sigma^2$. The new prior interpolates between the ICAR prior and the Gaussian prior for θ. λ serves as a spatial smoothing parameter and determines the proportion of the spatially structured risk variability over the total risk variability.

2.4.2 LLB model

The Leroux-Lei-Breslow model (LLB) model is based on a single set of random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, represented by a multivariate Gaussian distribution

$$\boldsymbol{\phi}|\mathbf{W}, \sigma^2, \rho, \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}). \tag{2.5}$$

The prior above has a constant non-zero mean $\boldsymbol{\mu} = (\mu, ..., \mu)$, avoiding the use of the intercept term in (2.1). In the matrix, $\sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}$, I_n is an $n \times n$ identity matrix and the elements of $\mathbf{W}^*$ are equal to

$$w^*_{ji} = \begin{cases} n_i, & \text{if } j = i \\ -1, & \text{if } j \sim i \\ 0, & \text{otherwise.} \end{cases}$$

The precision matrix is a weighted average of the spatially dependent correlation structures, represented by the matrix $\mathbf{W}^*$, the independent correlation structures, represented by the identity matrix, and the weight represented by the parameter ρ. When $\rho = 0$ the model becomes a simple independent random effects model and when $\rho = 1$ the model becomes the ICAR as in (2.2). When $0 \leqslant \rho < 1$ the joint distribution (2.5) is proper. The full conditional distributions corresponding to (2.5) are given by

$$\phi_i | \boldsymbol{\phi}_{-i}, \mathbf{W}, \sigma^2, \rho, \mu \sim N\left(\frac{\rho \sum_{j\sim i} \phi_j + (1-\rho)\mu}{n_i\rho + 1 - \rho}, \frac{\sigma^2}{n_i\rho + 1 - \rho}\right). \tag{2.6}$$

The conditional expectation is the weighted average of the random effects in the neighboring areas and the overall mean μ. The conditional variance, in the presence of strong spatial correlation is approximately σ^2/n_i, the same as the ICAR, but if the random effects are independent then it is a constant (σ^2).

2.4.3 Localized Conditional Autoregressive model (LCAR)

All three models defined above use priors that are globally smooth. The random effects are forced to exhibit a single global level of spatial smoothness determined only by geographical adjacency. With real data such a uniform level of smoothness for the entire region is unrealistic. It is more realistic to think that sub-areas of spatial autocorrelation co-exist with areas of discontinuity. As an example, areas of wealth and poverty, sharing boundaries, are very common in the biggest cities of the world, showing different patterns in the disease risk. A possible solution to this problem is presented by Lee, Rushworth and Sahu [52], and is called Bayesian Localized Conditional Autoregressive model (LCAR), LCAR from now on. This model was initially applied to a ESR, but as explained in Subsection 2.2.2 the same methodology can be applied in the DM field.

The LCAR treats the elements in the neighborhood matrix, representing contiguous areas, as a set of binary random quantities and not as fixed values. The elements of this new neighborhood matrix, $\tilde{\boldsymbol{W}}$, continue to be set to zero for non adjacent areas but adjacency is no longer the only reason for those elements to be set to one. When all adjacencies are kept, the model simplifies to the ICAR, while if all adjacencies are removed the random effects are independent. The model defines $\boldsymbol{\phi}$ in (2.2) as $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \phi_\circledast)$ where $\phi_\circledast$ is a global random effect that is potentially common to all areas and prevents any unit from having no information to "borrow strength" from. Based

on the extended matrix, the proposal is to model $\tilde{\boldsymbol{\phi}}$ as $\tilde{\boldsymbol{\phi}} \sim N(0, \sigma^2 \mathbf{Q}(\tilde{\mathbf{W}}, \epsilon)^{-1})$, with the precision matrix given by

$$\mathbf{Q}(\tilde{\boldsymbol{W}}, \epsilon) = diag(\tilde{\boldsymbol{W}}\boldsymbol{I}) - \tilde{\boldsymbol{W}} + \epsilon \boldsymbol{I}, \tag{2.7}$$

The component $diag(\tilde{\mathbf{W}}I) - \tilde{\mathbf{W}}$ corresponds to the ICAR model applied to the extended random effects vector $\tilde{\boldsymbol{\phi}}$ and the component ϵ ensures that the matrix is diagonally invertible. This restriction is now needed because $\mathbf{Q}(\tilde{\mathbf{W}})$ is no longer fixed. The parameter ϵ is recommended to be set as $\epsilon = 0.001$. The full conditional distributions corresponding to the LCAR model are given by

$$\phi_j | \boldsymbol{\phi}_{-j} \sim N\left(\frac{\sum_{i=1}^{n} w_{ij}\phi_i + w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon} \right) j = 1, ..., n \tag{2.8}$$

$$\phi_{\circledast} | \boldsymbol{\phi}_{-\circledast} \sim N\left(\frac{\sum_{i=1}^{n} w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon} \right).$$

In (2.8) the conditional expectation is a weighted average of the global random effect $\phi_{\circledast}$ and the random effects in the neighboring areas, with the binary weights depending on the current value of $\tilde{\mathbf{W}}$. The conditional variance is approximately (due to ϵ) inversely proportional to the number of neighbors remaining in the model, including the global random effect $\phi_{\circledast}$.

The matrix $\tilde{\mathbf{W}}$ is treated by the LCAR model as a single random quantity, which avoids several problems identified by other authors (for more details see Lee et al. [52], section 3.2). The authors [51] propose eliciting the set of candidate values of $\tilde{\mathbf{W}}$ from data having a similar spatial structure as the response variable.

The increased flexibility provided by the LCAR model inevitably means that it is more computationally demanding than the common BYM model.

2.5 Neighbourhood matrices

As mentioned above, DM uses areal data. There are three different types of data [5] when it comes to spatial data sets, point-referenced data, areal data and point pattern data:

1. *point-referenced data*, where $Y(s)$ is a random vector at a location $s \in \Re^r$, where s varies *continuously* over D, a fixed subset of $\Re^r$ that contains an r-dimensional rectangle of positive volume;

2. *areal data*, where D is again a fixed subset (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries;

3. *point pattern data*, where now D is itself random; its index set gives the locations of random events that are spatial point patterns. $Y(s)$ itself can simply equal one for all $s \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a marked point pattern process).

Point-referenced data are also called geocoded or geostatistical data, and are built of specific locations of, for example, air-pollution sites, hospitals, etc.. The classical approach to spatial prediction in the point-referenced data setting is *kriging*. An interesting comparison between Poisson kriging and the BYM model can be found in Goovaerts and Gebreab [31].

An example of the third type of data can be residences of people suffering from a particular disease, and the questions of interest typically centre on whether the data are *clustered* more or less than would be expected if locations were determined completely by chance. Statistics that measure clustering are often use in this context, with the most common being *Ripley's K function*. More details can be found in Diggle's book [23].

Most areal data are summaries over an *irregular* lattice, like a collection of counties or districts. As mentioned by Banerjee et al. [5], the primary concept in the initial exploration of areal unit data is the *proximity matrix*. Below we detail the several types of matrices actually used in the literature.

2.5.1 Adjacency matrices

During the description of the DM models (see section 2.4) we have already introduced one type of neighbourhood matrix, the one that is most widely used, the adjacency matrix. Based on the Markovian property, this matrix assumes conditional independence beyond space neighbourhood.

We presented three different matrices, the one used in the BYM model, the one used in the LLB model, and the one used in the LCAR model (see previous section for the matrix elements definition).

Adjacency matrices can be classified regarding the:

a. order, adjacency matrices can be of first-order, or *"nearest-neighbour"*, second-order, the *"neighbours of the neighbours"*, and so on to the n^{th}-order.

b. Rook or Queen, adjacency matrices can be based on common points, using only boundaries (Rook), or using both boundaries and vertices (Queen).

c. basis, adjacency matrices can define neighbourhood based on administrative zones (created by authorities to meet goals that have little to do with disease etiology or common risk factors) or by redefining the full area into small areas or grids.

d. fixed or estimated, adjacency matrices can be fixed, or random quantities resulting from some kind of statistical process.

Examples of the above categories of matrices can be found in the literature. For the order and Rook or Queen classification categories, Earnest et al. [24] use four different matrices, using first- and second-order contiguity, boundaries only (Rook), and boundaries and vertices (Queen). For the basis classification category a good

example is presented in English et al. [25] in which the area was redefined in grids of 0.5 miles and neighbours chosen were the immediate north, south, east, and west of each particular gridpoint. A good example of the estimated adjacency matrix is the one used in the LCAR model [52].

As Raftery and Banfield [69] point out in the discussion of Besag et al. [9] seminal work, whilst the adjacency definition seems to be acceptable when the partition of areas is not too dissimilar to a regular array of pixels (reflecting the original use of such models in image analysis problems), it may be less appropriate for the much more irregular spatial partitions found in DM applications.

2.5.2 Distance matrices

Another type of neighbourhood matrix is the distance matrix, created mostly with the goal of overcoming the above-mentioned problem of the irregularity in the areas. In this matrix definition the conditional dependency between two areas decreases quickly as the distance in space increases. Several types of specifications have been used, usually all parametric functions of distance [11, 19].

Also, here the concept of order can be introduced, by creating distance bins, say $(0, d_1], (d_1, d_2], (d_2, d_3]$ and so on, enabling the notion of first-order neighbours of unit i, meaning all units within distance d_1 of i, second-order neighbours, meaning all units more than d_1 but at most d_2 from i, and so on.

We will detail here the distance matrix defined in Best et al. [11], as this is the matrix that will be used later in this work. Defining the matrix by $\mathbf{D}$, with elements $d_{ij} = d_{ji}$, where

$$d_{ij} = e^{-k_{ij}/\delta}$$

for k_{ij} = distance in kilometres between the geographic centroids (the Euclidean distance) of district i and j. Authors defined δ as a fixed value, arbitrarily chosen to give a relative weight of 1% to an area j, whose centroid has the mean inter-district distance for the full area from area i.

To find matrices based on non-Euclidean distances, we need to leave DM and go into the geographical sciences and spatial econometrics spaces. Geographically weighted regression [26] is an important local technique for exploring spatial heterogeneity in data relationships. Geographically weighted regression makes a point-wise calibration concerning a "bump of influence": around each regression point where nearer observations have more influence in estimating the local set of coefficients than observations farther away. In the estimation of the regression coefficients a diagonal matrix denoting the geographical weighting of each observed data for regression point i is used. While geographically weighted methods deal with spatial non-stationarity, spatially autoregressive methods [3] deal with spatial dependence. Both are a function of many factors, including the nature of the phenomena under investigation and the representation of space underpinning the model. One of the most crucial parameters

on that representation, affecting the analytical results, is the distance measurement method.

A geographical space is usually too complicated to be measured simply by Euclidean metrics (to build the neighbourhood matrix), which may fail to reflect true spatial proximity and instead, non-Euclidean metrics should be used, such as road network distance, travel time, water distance or landscape distance.

One example of a geographically weighted regression with a non-Euclidean distance metric is presented by Lu et al. [56]. A case study, a London house price data set coupled with hedonic independent variables, where models are calibrated with Euclidean distance, road network distance and travel time metrics. Authors conclude that non-Euclidean distances improve model fit, and provide additional and useful insights.

One example of a spatial autoregressive model using different types of distance estimation is provided by Shahid et al. [78] on a study in spatial analytical modelling for health care planning. Euclidean, Manhattan and Minkowski distance metrics are used to estimate distances from patient residence to hospital. Distances estimated with each metric are contrasted with road distance and travel time measurements. Authors conclude that the Minkowski method produces more reliable results than the traditional Euclidean metric. Formally, the Euclidean, Manhattan, and Minkowski distance can be calculated by the formula:

$$d = \left[(x_i - x_j)^p + (y_i - y_j)^p\right]^{1/p} \tag{2.9}$$

where, x and y are the geographical coordinates of the centroids of point i and point j. The generic p in equation 2.9 is replaced by the value 2 to yield Euclidean distance; the value 1 would yield the Manhattan distance, and all intermediate values in the $[1 < p < 2]$ interval yield an array of Minkowski distances.

2.5.3 Measures of spatial association

Finally, we introduce here the two standard statistics used to measure the spatial association among areal units [74], Moran's I and Geary's C.

Moran's I takes the form

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}, \tag{2.10}$$

and Geary's C takes the form

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}. \tag{2.11}$$

I is not strictly supported on the interval $[-1, 1]$ and is a ratio of quadratic form in $\mathbf{Y}$. Under the null model where Y_i are independent and identically distributed, I is

asymptotically normally distributed with mean $-1/(n-1)$ and variance

$$Var(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}. \tag{2.12}$$

In equation 2.12, $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = 1/2 \sum_{i \neq j}(w_{ij} + w_{ji})^2$, and $S_2 = \sum_k (\sum_j w_{kj} + \sum_i w_{ik})^2$.

C is never negative, has a mean of one for the null model, and values $[0, 1]$ indicate positive spatial association.

To run a true significance test using equation 2.10 or equation 2.11 a Monte Carlo approach should be used. Under the null model the distribution of I (or C) is invariant to permutations of $Y_i's$. A Monte Carlo sample of, say 1 000 permutations, including the observed one, will position the observed I (or C) relative to the remaining 999, to determine whether it is extreme, via an empirical p-value.

2.6 Concluding remarks

In the past years hierarchical Bayesian models have been developed and refined to achieve statistically precise local estimates of disease risk for each small region. In this chapter four of those models are presented:

1. The BYM model as the most widely used.

2. The MBYM model which derives from the BYM model in an attempt to overcome the known deficiency of the latter, the lack of identifiability. The MBYM is identifiable and facilitates hierarchical modeling of the additive effects of unobserved covariates that might be spatially and randomly varying [59]).

3. The LLB model as the one that has consistently shown [47, 59] good results across a variety of cases.

4. The LCAR [52] model as the only model that does not take the neighborhoods as fixed but those emerge from real data, as a random quantity.

There are still many opportunities for future work in this area. One of the most evident is the global ICAR's property of tending to negative pair-wise risk dependance as the "spatial proximity" between two regions is further apart and its potential impact on posterior inference has not been yet sufficiently explored and understood.

We also presented the neighbourhood matrices types actually used in DM. There has not been a great deal of work dedicated to this subject, and the types of matrices available are therefore restricted. Due to the type of data used, areal data, only spatial adjacency or distance based matrices exist. The neighbourhood matrix provides the mechanism for introducing spatial structure into the formal modelling. The elements of the matrix can also be seen as weights, with more weight being associated to $j's$ closer (in some sense) to i than those further away from i.

Chapter 3

Data

3.1 Introduction

The contents of this chapter is based on the paper: **Implementing the World Mental Health Survey Initiative in Portugal - rationale, design and fieldwork procedures [94]** and in the paper: **Alcohol abuse disorder prevalence and its distribution across Portugal. A disease mapping approach [7]**.

Recent epidemiological research shows that psychiatric disorders and other mental health-related problems have become the main cause of disability, and one of the main causes of morbidity and premature death throughout the world [67].

Mental disorders are responsible for more than 12% of the global burden of disease in the world as a whole – a figure that rises to 23% in developed countries. Five of the 10 main causes of long-term disability and dependency are neuropsychiatric conditions: unipolar depression (11.8%), alcohol-use disorders (3.3%), schizophrenia (2.8%), bipolar disorders (2.4%) and dementia (1.6%) [92]. In Europe, mental health problems account for nearly 26.6% of the total burden of ill health, while suicide is one of the top ten leading causes of premature death [72]. Estimates from the European Brain Council indicate that 27.4% of the EU population aged 18 to 65 suffer from one type or another of mental health problem during each one-year period [89], a number that has been recently updated to 38.2% after the inclusion of data from a broader childhood and adolescence assessment, as well as from new EU member states [90].

While there are people who have a diagnosable disorder, many others have mental health problems that can be considered "subliminal", meaning that they do not meet the diagnostic criteria for psychiatric disorders, but are also in distress, and should therefore benefit from intervention.

Furthermore, people with mental health problems are more likely to have physical health problems, leading to a significant impact on family life, social networking, job performance, and employment, as well as to suffer from stigma, discrimination, and social exclusion. In fact, there is reliable evidence that in several places basic human rights may be denied to people with mental health problems [43]. Besides the burden of disease associated with psychiatric conditions, economic impact should also be taken into account: for instance, in the United Kingdom alone, the cost to the economy (direct health costs, welfare benefits, lost productivity at work) was estimated at over £77 billion every year [64].

In the last two decades, psychiatric epidemiological studies provided a relevant contribution to unveil the dimension, determinants, social impact and treatment gap of the psychiatric disorders.

Thanks to the refinement of survey methodology and questionnaire development, it has been possible to carry out studies based on fully-structured interviews of large samples of the general population, administered by trained lay-interviewers, such as the Epidemiological Catchment Area Study [71] and the National Comorbidity Survey [40], which showed rates of prevalence of psychiatric disorders close to 30% in the year prior to the interview.

Ten years ago, the World Health Organization (WHO) and Harvard University jointly decided to promote a worldwide initiative of population-based surveys using the same methodologies – the WMHSI [22] – aiming to improve the knowledge on the natural history, magnitude and impact of mental illnesses. Conducted in more than 30 countries worldwide (in the Americas, Africa, Europe, Western Pacific and South-East Asia), with a total sample size that can exceed 154 000 people, this project has provided so far, through more than 500 published papers, a huge amount of crucial epidemiological information regarding the planning and implementation of mental health policies (for further details please refer to the project's website, available at: http://www.hcp.med.harvard.edu/wmh).

In 2007, after the launching of a new national mental health plan, it was decided to carry out a national survey in Portugal, in order to overcome the scarcity of representative data about the prevalence of psychiatric disorders and mental health problems in the country. Grounded on the data from the European Brain Council Report entitled "Costs of Disorders of the Brain in Europe", it has been indirectly estimated that 1 557 054 (16.07% of the adult population - 18 to 65 years) have a mental disorder in Portugal. According to this projection, 5.09% of the Portuguese adult population suffer from affective disorders (including dysthimia), 9.46% from anxiety disorders, and 0.52% from psychotic disorders [89]. Beyond this projection, general psychological morbidity data from the Eurobarometer survey has suggested that the prevalence of mental health problems in Portugal could be higher than in other European countries of similar characteristics, while the most vulnerable groups (women, the poor, the

aged) seemed to exhibit also a higher risk of psychiatric caseness than in the rest of Europe [77]. Several other studies, although conducted with non-representative samples, seem to point in the same direction [42, 44, 62].

Despite the available data suggesting the existence of significant levels of psychiatric morbidity and unmet needs for care throughout the Country, there are still no sound published figures about the prevalence of psychiatric disorders in Portugal. To collect comprehensive and representative epidemiological data, a survey following the methodological framework of the WMHSI was carried out in Portugal, between 2008 and 2009. This study, overseen by the WHO and Harvard University, was coordinated by the Nova Medical School (Department of Mental Health, NOVA University of Lisbon). This is the first rigorous general population survey carried out with a nationally representative sample of the Portuguese population, aiming to evaluate the prevalence, the correlates, the impact and the treatment patterns of mental disorders. This chapter presents an overview of the methodology implemented in Portugal, covering the main features of the study (design, fieldwork organization, sampling and weighting procedures).

In Section 3.2 we look in depth to the methods used to collect the data. Section 3.3 presents some considerations on the challenges faced by the implementation team and the importance of the study. In Sections 3.4 and 3.5 some results regarding AAD are presented, using several different methodologies but special attention is dedicated to the results of the models reviewed in Section 2.4.

3.2 Methods

3.2.1 Design and general framework

Following the original methodology designed by the WMHSI [22], the Portuguese mental health survey is a cross-sectional study based on stratified multistage clustered area probability household sample. It was carried out at the households of a Portuguese nationally representative sample of respondents, between October 2008 and December 2009. The survey was administered by trained lay-interviewers on a face-to-face setting, using the Computer-assisted personal interview (CAPI) methodology. The use of CAPI rather than the Paper-and-pencil assisted personal interview (PAPI) version was a decision from the WMHSI Data Collection Coordination Centre (Harvard University), in order to avoid problems related with data input costs, increased length of interview and dropping-out, as well as to facilitate quality control. Given the complexity of the sampling procedures and the fieldwork, a highly specialized survey unit, Center for Public Opinion Studies and Polls (CESOP) belonging to the Portuguese Catholic University, was selected to implement the protocol throughout the Country, under the scientific coordination of the Nova Medical School. The project was submitted to and approved by the Ethics Committee of the Nova Medical School in January 2008.

3.2.2 Target population

The National Census 2001, published by Statistics Portugal - Instituto Nacional de Estatística (INE) [37], was used to estimate the target population of mainland Portugal in 2008. The target population for the survey was defined as the usually resident, non-institutionalized Portuguese-speaking population of Continental Portugal aged 18 or above, residing in permanent private dwellings. This definition excluded: a) People living in non-private dwellings, b) Residents of rest homes, hospitals and psychiatric institutions, c) Military personnel not residing in a private dwelling, d) Prison inmates, e) Non-Portuguese speakers and f) Other people unable to answer the questionnaire. Recent data from the National Census 2011 [38] confirms that the population that meet the criteria a) to d) represents 1.3% of the total population aged 20 years or more; the population that meet the criteria e) is unknown, but Census 2011 estimates it at around the 1% level. Anyway, two limitations have remained: 1. the above percentage of 2.3% (1.3% + 1.0%) was measured in 2011 and not in 2008, although it is not expected that major changes had occurred between 2008 and 2011 and 2. the size of the population belonging to criteria f) is not known.

3.2.3 Sampling

This is a stratified four-stage clustered area probability design, using the "locality" as the Primary sampling unit (PSU). Unlike what happens in several countries, which have reliable lists of residents available for survey research – such as Sweden, for example – or where there are reliable lists of households/addresses – such as the UK, for example – in Portugal no such lists are available. Therefore, a multistage design needs to be applied, in which the selection of localities forms the first stage. "Localities" are territorial delimitations defined in the context of census operations by the INE, consisting in population clusters with 10 or more residential dwellings and to which a distinct place name is attached. For each locality, the number of households and people 15 years and older is known on the basis of census data. According to the Census 2001, 27 960 localities existed in Portugal mainland, with 7 719 986 inhabitants aged 18 or more. At stage 1, 262 PSU were randomly selected with probability proportional to size, as shown inTable 3.1. Selection was stratified by region (North, Centre, Lisbon, Alentejo, and Algarve) and size of locality (1- $\leq$ 2 000 inhabitants; 2- 2 000 - 9 999 inhabitants; 3- 10 000 – 19 999 inhabitants; 4- 20 000 – 99 999 inhabitants; 5- $\geq$ 100000 inhabitants). The number of non-empty strata created was of 23 (there were no localities with 100 000 inhabitants or more in the regions of Alentejo and Algarve).

Out of the 262 PSUs selected, 52 PSUs entered the sample with certainty. The certainty selection included all localities with population larger than 20 000 inhabitants. Those 52 PSUs are referred to as "self-representing" PSUs because they were not selected randomly to represent other localities, but they are so large that they represent themselves. The remaining ones are "non-self representative" because they were

Region	Size of number of inhabitants in the localities PSU				
	≤ 2000	2 000 - 9 999	10 000 - 19 999	20 000 - 99 999	≥ 100 000
North	53	15	9	20	2
Center	48	9	5	6	1
Lisbon	12	18	11	17	2
Alentejo	10	8	2	2	
Algarve	6	2	2	2	
Total	262				

Table 3.1: Number of localities randomly selected (stratified by region and locality size).

Region	Size of number of inhabitants in the localities PSU				
	≤ 2000	2 000 - 9 999	10 000 - 19 999	20 000 - 99 999	≥ 100 000
North	54	9	16	19	2
Center	46	4	9	6	1
Lisbon	12	11	17	15	2
Alentejo	11	1	8	2	
Algarve	6	2	1	2	
Total	256				

Table 3.2: Number of localities (stratified by region and locality size) where interviews were conducted.

selected to be representative of smaller areas of the country. In the end, the PSUs with interviews were distributed as shown in Table 3.2.

At stage 2, there was a selection of random-route starting points. By means of aerial maps, coordinates were randomly selected. Initially, for PSUs with less than 100 000 inhabitants, a total of 4 to 6 routes starting points were selected. For the "self-representing" PSUs above 100 000 inhabitants a number between 12 and 43 starting points were selected.

At stage 3, the initial selection of households was conducted. Using the already mentioned 2001 Census information on the number of households in each locality, households were selected by applying intervals proportional to size of locality and divided by number of random-route points to be selected in each locality. This design creates a sample in which the probability of any individual Housing Unit (HU) being selected to participate in the survey is equal to every HU in Portugal mainland. A total of 10 067 addresses were selected.

At stage 4, information regarding the number of people aged 18 years old or more living in the household was collected. The interviewers registered each selected address and, if someone was present, registered gender and birth date of qualified members of the population in each household, leaving information about survey and collecting phone numbers. Following analysis of interviewer records and validation of

choice of household and/or respondent by CESOP central coordinators, based on the last birthday method (in which an interview is attempted with the adult in the household who had the most recent birthday), 8 253 respondents were selected, under the expectation of a 50% cooperation rate.

3.2.4 Tools and measures

The interview tool is the World Mental Health - Composite International Diagnoses Interview (WMH-CIDI), a new expanded version of the WHO-Composite International Diagnoses Interview (CIDI), developed by the WMHSI and the National Institute for Mental Health [41]. The original CIDI is a fully structured questionnaire on the presence, persistence and intensity of clusters of psychiatric symptoms and provides, by means of computerized algorithms, lifetime and 12-month mental disorders diagnoses according to the International Classification of Diseases 10^{th} Edition (ICD-10) [93] and accordingly with the Diagnostic and Statistical Manual of Mental Disorders 4^{th} Edition (DSM - IV) [2]. The WMH-CIDI was developed in order to overcome detected shortcomings of the original version, and was extended to include accurate questions about disorder severity, impairment, and treatment. Kessler et al. [41] and Haro et al. [33] provided evidence that diagnoses of anxiety, mood, and substance disorders based on CIDI 3.0 have generally good concordance with diagnoses based on blinded clinical reappraisal interviews. The adaptation of the original WMH-CIDI to Portuguese was conducted by a committee from the Nova Medical School, including 10 bilingual experts with clinical experience, coordinated by two of the study responsibles (José Caldas-de-Almeida (JCA) and Miguel Xaxier (MX)), in close contact with the WMHSI Data Collection Coordination Centre. The process was driven according to 5 specific dimensions: semantic equivalence (likeness of meaning of each item), content equivalence (item cultural relevance), technical equivalence (use of the same measuring techniques), conceptual equivalence (relationship of the theoretical constructs against criteria known to be related) and criterion equivalence (similarity of the results of the measure in the two cultures). To achieve the Portuguese final version, a step-by-step procedure was used as detailed below.

1. The process started with a general review of all sections, aiming at the identification of words, phrases and idiomatic expressions not commonly used in Portugal, and listing of suggested language equivalents.

2. A specialized review of each section was conducted by two members of the committee, according to their specific areas of expertise. The reviewers were asked to continue the identification of words, phrases and idiomatic expressions not commonly used in Portugal, and to make recommendations for modifications to the Portuguese version, aiming for equivalence (semantic, content, conceptual, and technical) with the original English version.

3. The instrument was then reviewed item by item by a group (JCA, MX and the people responsible for the second review) focusing on the modifications made to fit the vernacular use in Portugal and working for consensus on items that were thought to be problematic.

4. All sections and items were crosschecked by the same group for consistency in word use throughout the instrument.

5. A new version was created with all of the agreed upon cultural adaptations.

6. The entire newly adapted instrument was administered to 71 potential respondents of the target population to test for respondents' reactions and understanding of particular questions that people do not seem to understand or seek clarifications on.

7. Last changes and confirmation of the final version by JCA and MX, after results from the pilot test.

Disorders considered in the Portuguese version of the WMH-CIDI include anxiety disorders (agoraphobia, generalized anxiety disorder, obsessive-compulsive disorder, panic disorder, posttraumatic stress disorder, social phobia, specific phobia), mood disorders (bipolar I and II disorders, dysthymia, major depressive disorder), disorders that share a feature of problems with impulse control (bulimia, intermittent explosive disorder - for all respondents - and adult persistence of childhood/adolescent disorders- attention deficit/hyperactivity disorder, conduct disorder, and oppositional-defiant disorder - only for respondents in the 18 to 44-year age range), and substance disorders (alcohol abuse and dependence).

In addition, the instrument includes modules intended to assess, amongst others, various areas of life, including marriage, work, financial issues and education. Other modules included are the functioning and physical disorders (stigma, discrimination, number of days unable to work, diseases severity, childhood adversities and suicide) and the treatment (psychiatric treatment, other mental health treatment, any professional treatment, any health treatment, and reasons for no treatment).

Once prepared, the content of the Portuguese version was transferred to a CAPI using the Blaise software [80], a task that involved reprogramming the codes received from the WMHSI Data Collection Coordination Centre. Special attention was given to the skips and jumps within and between sections, in order to fully respect the original CAPI algorithms.

The instrument was pre-tested in 71 respondents living in the district of Lisbon. The distribution of the people on the sample was based in the distribution of the Portuguese population presented in the Census 2001, thus representing all adult age groups, males and females, from different socio-economic strata. Interviews were

conducted by 10 previously trained lay-interviewers, appointed by the CESOP survey centre to carry on the supervision of the field work.

The data from the pilot study was sent to the WMHSI Data Collection Coordination Centre on May 2008. During the data cleaning some skip errors were found both within and between sections, which led to several amendments introduced in the final version of the Portuguese WMH-CIDI CAPI.

Besides minor problems related with some questions considered as being confuse, repetitive and too long, the main problem found was that the complete interview could take up to 180 minutes. Internal sub sampling was used to reduce respondent burden by dividing the questionnaire into two parts. During the field work, all respondents completed Part I, which included screening questions and assessed core mental disorders. All respondents that met the criteria for any DSM - IV disorder were then administrated Part II, the diagnostic, additional disorders and correlates modules. Beyond that, Part II was also administrated to a probability sample of 25% randomly selected by the Blaise Code of those who did not meet criteria for any disorder. Therefore, to out of a total of 3 849 interviews, 2 060 were administrated the "long interview" (Part I + Part II) and the remaining ones were administrated the "short interview" (Part I).

3.2.5 Fieldwork organization and procedures

3.2.5.1 Procedures

After sample selection, each selected household received a brochure asking for participation in the study, presenting the objectives, the work team and the free phone number to contact the supervisors' team. An introduction letter, signed by the Principal Investigator, was sent at the same time. At the first visit, and after confirming availability, household was centrally confirmed and a phone contact was undertaken to schedule interview. If availability was not confirmed either new visits were conducted or a refusal was registered (see below, for call procedures detailing fieldwork guidelines regarding household visiting, interview scheduling and substitutions).

3.2.6 Call procedures (fieldwork guidelines regarding household visiting, interview scheduling and substitutions)

a. If the household report available after first visit and selection of household is centrally confirmed, contact by phone is made to schedule interview;

b. If household report not available after first visit, subsequent visit is made to fill household report; number of visits before household deemed "unknown if occupied": 9;

Sample distribution	%	(n)	%
Interview	38.2	3 849	57.3
Refusal	28.5	2 865	42.7
Sub-total		6 714	
No contact	14.0	1 408	
Circunstancial (outside the sample definition)	17.0	1 707	
Others	2.3	238	
Total		10 067	

Table 3.3: Sample Distribution.

c. Number of established contacts with household with failure to contact selected respondent before deemed "non-contact": 5 phone contacts, and 4 household visits before deemed "non-contact";

d. Number of scheduled interviews with failure to attend by respondent before being deemed "refusal": 2;

e. No substitutions: all respondents are extracted from initially selected households.

Table 3.3 shows the final sample disposition. From the total 10 067 households selected, 3 849 participated in the survey, representing 57.3% of total contacted (6 714).

The response rate achieved was close of that achieved in Belgium, France, Germany, and Netherlands, but lower than that achieved on others countries, like the USA [13]. Informed consent was asked and obtained in every occasion, by means of a signed form, previously accepted by the Nova Medical School Ethics Committee.

3.2.7 Data management and quality control

Once surveys were completed, data was sent to the World Mental Health (WMH) Data Analysis Coordination Centre at the Department of Health Care Policy, Harvard Medical School, for data cleaning. The first data file of the study was sent in July 2008, containing 711 completed interviews. The cleaning process revealed minor problems concerning a small number of CAPI outputs, namely due to incorrect skips between questions/sections of the CIDI, that were easily corrected.

Once cleaning was completed, centralized coding and analysis were carried out. Cleaned and coded data sets and the results of preliminary analysis were sent back to the Portuguese team.

A comprehensive system of quality assurance was set from the starting of the field-work, including several simultaneous mechanisms. Appointments were monitored by random call-backs to households (10% of total) using part of the stem questions in the

"Screening" section, as well as other questions created by the management team (ex. *"Did you receive the incentive? How much was it?"*). During these random call-backs, supervisors evaluated i. study eligibility i.e. if the appropriate respondent was correctly interviewed (instead of another household resident), ii. response to key questionnaire items, iii. date, time and total length of the interview (to cross-check with the duration displayed by the software program), iv. respondent's feedback on interviewer's professionalism and v. compliance with the interviewing rules and guidelines set forth in the training. The WMHSI Coordinating Centre provided a software that use the clocks in the laptops to time data entry as a way to detect possible interviewer cheating, also allowing the rapid analysis of computerized CAPI interviews to detect missing values and other signs of low interview quality. A deliberate violation of interview guidelines was found in one occasion, leading to immediate exclusion of the interviewer. Besides the indirect evaluation, fieldwork supervisors from the CESOP directly observed up to 10% of all interviewers' work, selected randomly from the case register. Monitoring was more frequent earlier in the study, but to ensure reliability supervisors continued to monitor even the experienced interviewers till the end of the study, in addition to the less experienced interviewers.

3.2.8 Weighting

When conducting a survey, having a representative sample of the population is of paramount importance, but despite the best efforts some characteristics (such as age, education, race, gender, etc.) of the sample might be accidentally oversampled or under-sampled. Correction can be accomplished via a post-stratification.

On the present data, two different weightings were considered. Weight on total sample (WT1) will be used when the total sample (n=3 849) is considered, while Weight on reduced sample (WT2) will be used just for the respondents answering the long interview (n=2 060).

WT1 weighting was calculated based on two different weights. Firstly, the number of eligible respondents in each household was computed, allowing to the calculation of the within-household weight (weight 1 of WT1). The within-household probability of selection weight adjusts for the fact that the probability of selection of respondents within the HU varies inversely with the number of people in the HU. This is true because, as noted earlier, only one respondent was selected for interview in each HU. Three variables in the household listing were included at the individual-level records: respondent age and gender and number of eligible residents in the HU. If the number of eligible respondents in the household was greater than 5, then 5 was used. This was the weight 1 of the WT1.

To compute the second weight (weight 2 of WT1), and adjust for variation between the joint distribution of age-gender in this weighted sample compared to the INE published data, a post-stratification by those two socio-demographic variables for each

Gender/Age	Unweighted %	INE
Gender		
Male	42.4	48.3
Female	57.6	51.7
Age		
18-34	27.2	29.4
35-49	30.3	28.3
50-64	25.5	23.6
+65	17.0	18.7

Table 3.4: Unweighted sample composition and INE published data.

of the 5 regions of Portugal was done.

WT1 is equal to weight 1 × weight 2. The sum of the WT1 is now the population size of the 5 regions in Portugal mainland, aged 18 or more, as it was published by INE, as the annual Portuguese resident population for 2008, in September 15, 2009. This weight was then normalized to the sample in order to obtain the final WT1. After that, the upper and lower 3% of cases, in the final WT1, were trimmed by assigning them the average value of the total weight and therefore obtaining the "weight trim". After this, the "weight trim" becomes the final WT1.

Comparison of the unweighted distributions of the sample with the INE published data distributions provides information justifying the weighting above described. As Table 3.4 shows, the sample overrepresented women and people with 35–64 years of age – these distortions were corrected with the WT1 weight.

WT2 weighting was also calculated based on two different weights. Firstly, a group was assign to each respondent, "core disorders" or "no core disorders". 572 respondents without core disorder got the long interview – in fact, they represent a total of 2 342 respondents that did not have a core disorder (572 without core disorder that got the long interview + 1 770 without core disorders that did not got the long interview). Therefore, weight1 of WT2 is "1" for the interviews that have a core disorder and "572/2 342" for the ones who do not have a core disorder. After multiplying those values by WT1, the final weight 1 of WT2 was obtained.

The weight 2 of WT2 was calculated exactly the same way as weight 2 of WT1 (based on age and gender, on regional level). WT2 was firstly calculated multiplying weight 1 × weight 2, followed by normalization of the interview sample of 2 060 (long-interviews).

Depending if a disorder is included in the long or short interview, WT1 or WT2 will be used accordingly in order to calculate the prevalence of each disorder.

3.3 Challenges and Importance of the study

The implementation of the WMH survey in Portugal was a difficult process, but also a very important challenge for the teams involved in the project. Generally, the type of problems and limitations faced during the implementation were rather similar to the ones already described by other countries belonging to the WMHSI [68]. Firstly, this is a very demanding, time-consuming and costly project, with methodological requirements highly specific and quite demanding. Cost-efectiveness has been pointed, in fact, as one of the major problems of the WMH surveys, with some authors arguing that these cross-sectional studies might not be cost-effective [86], once relying in potentially biased retrospective reports. Secondly, it is plausible that people with serious mental disorders could be more prone to refuse being interviewed (i.e., systematic non-response), which could lead to bias regarding the estimation of disorder prevalence. The same applies to non-reporting, an issue that is often relevant in very lengthy interviews, in which the respondents may get tired or bored. Looking from a more positive perspective, the relatively small size of the country and the quality of the data provided by Statistics Portugal were undoubtedly factors that facilitated the fieldwork. As well, group cohesion and motivation ensured by the CESOP were crucial ingredients to the success of the study, avoiding a high turn-over of interviewers and thus decreasing the need for further CIDI training during the project. The coordination team was in close contact with Harvard University during the whole study, which allowed an excellent scientific support in the data management process, namely on data cleaning. Two of study coordinators (JCA and MX) have participated in the Annual WMHSI Meetings since 2008, presenting updated data on the implementation of the study in Portugal.

The results of this study may be of great importance to the Portuguese authorities, as it will provide the Government with the first nationally representative data on the magnitude and distribution of the mental health disorders in Portugal. In fact, despite some indubitably positive aspects, due to lack of planning and consistent support in the improvement of mental health services, Portugal is still lagging behind in this field in relation to other European countries. Existing data and analysis of results from research undertaken as part of the National Mental Health Plan Report show that mental health services suffer from serious deficiencies, in terms of accessibility, equity and quality of care. Many local mental health services continue to be limited to hospitalization, outpatient consultations and, sometimes, day hospital, and have no community mental healthcare teams, with integrated case management, crisis intervention and programs involving families. On the other side, the number of people in contact with public services shows that only a small part of those with mental health problems have access to specialized mental health services. Even assuming that only people with severe mental illnesses attend mental health services – which we know is not the case – the number of contacts (17% of the population) is still extremely low in relation to what should be expected.

This study will provide currently unavailable sound data on patterns and correlates of service use and barriers to obtaining available treatment.

This is the first general population survey of psychiatric morbidity conducted in a nationally representative sample of the Portuguese population. The findings of this study can have a major influence in mental health care policy planning efforts over the next years, specially in a country that still has a significant level of unmet needs regarding mental health services organization, delivery of care and epidemiological research.

3.4 Alcohol Abuse Disorder - Exploratory results

This is a cross-sectional study, meaning that both disease cases and possible risk factors are collected at the same time. As reported in Waller and Gotway [85] that restricts the conclusions that can be drawn from the models. It is not possible to establish causal relationships between disease cases and possible covariates.

Data collected by cross-sectional studies may have several types of biases, as already pointed. Other types of biases can be: a) the possibility of selection bias, as only the non-institutionalized population and the population above 18 years of age was selected, and accordingly to the WHO [88] the alcohol consumption is rising between adolescents (13-18 years of age) and young adults. Therefore inferences can only be made on the study population and not on the global Portuguese population; b) another possible common bias is the misclassification bias, *i.e.*, the incorrect assignment of a disease to the study participants. This type of bias may occur in studies like this one, because there is no intervention of a medical doctor during the questionnaire's self-administration. As already mentioned, this problem seems to have been solved in France, Italy, Spain and United States of America, as Kessler et al. [41] and Haro et al. [33] provide evidence that the diagnoses of substance abuse disorders identified by the questionnaire used in this initiative, the CIDI 3.0 have generally good concordance with diagnoses based on blinded clinical reappraisal interviews. Unfortunately those tests have not been conducted in Portugal. Although the alcohol consumption and related disorders are very much connected with cultural aspects [87], we think that the performance while identifying the actual presence of disease has not been seriously affected.

3.4.1 Prevalence

According to the WHO [88] approximately 5.1% of the global burden of disease, and 5.9% of all deaths worldwide are attributable to alcohol consumption. Furthermore, harmful use of alcohol inflicts significant social and economic losses on individuals and society at large.

In accordance with the DSM - IV [2] criteria there are two possible diagnoses of alcohol disorders, the alcohol abuse disorder and the alcohol dependence disorder. In the six European countries [1] covered by the European Study of the Epidemiology of Mental Disorders (ESEMeD) project[1], 5.2% of the respondents report a lifetime history of alcohol abuse and/or dependence disorders. From the data collected in Portugal the prevalence rate of a lifetime history of alcohol abuse and/or dependence disorders is 10.0%, while the last 12-month prevalence rate is 1.6%; the lifetime prevalence rate of alcohol dependence disorder is 1.3%, while the last 12-month prevalence rate is 0.26%; the lifetime prevalence rate of alcohol abuse disorder is 8.7%, whereas the last 12-month prevalence rate is 1.3%. The high prevalence of alcohol abuse disorder found in Portugal reiterates the need to maintain alcohol abuse as a public health priority in the country, and therefore more detailed studies are needed.

3.4.2 Individual level covariates

Structured additive regression models, which among others include the Generalized linear model (GLM) and the Generalized additive model (GAM), are perhaps the most commonly used class of models in statistical applications [76]. The linear model, "created" in the XIX century by Legendre and Gauss, has been continuously developed since then. These developments intend to help explaining phenomena inadequately explained before. Those developments led to the GLM, for which the comprehensive reference is McCullagh and Nelder [61], and to the GAM, as defined by Hastie and Tibshirani [34]. The theory developed in this subsection is based mostly on the work of Wood [91].

A GLM model is defined as:

$$g(\mu_i) = \mathbf{X}\boldsymbol{\beta}, \tag{3.1}$$

Nowadays, the tendency has been to move away from the linear functions and model the dependance of the response variable $\mathbf{y_i}$, a vector with $i = 1, \ldots, n$ observations on random variable Y, on the covariates in a nonparametric fashion [34]. A GAM is a GLM in which a part of the linear predictor is specified in terms of smooth functions ($\mathbf{S}$) of the covariates $\mathbf{X}_i$[91]. This adds some complexity to the model, because the smooth functions need to be represented and the degree of smoothness needs to be determined. Formally a GAM is:

$$g(\mu_i) = \mathbf{P}_i\boldsymbol{\beta} + \mathbf{S}(\mathbf{X}_i), \tag{3.2}$$

where $\mu_i = \mathrm{E}(Y_i)$, g is a canonical "link function", β is a vector of unknown parameters, $\mathbf{P}_i$ is the i^{th} row of the model matrix for only parametric model components, and $\mathbf{X}$ is a $n \times q$ model matrix of q non-parametric components, with $\mathbf{P}$ and $\mathbf{X}$ representing

[1] The ESEMeD Project was created to fully study the results of the WMHSI on the following countries: Belgium, France, Germany, Italy, the Netherlands and Spain. As Portugal joined the WMHSI later than others, most of the publications, including the above mentioned [1], do not include Portuguese results.

the *predictor variables*. The link, g, is the function such that $g(\mu_i) = \theta_i$, where θ_i is the canonical parameter of the distribution.

The GLM allows that the response variable Y follows any distribution among those belonging to the exponential family. A distribution belongs to the exponential family of distributions if its probability density (or mass) function, can be written as

$$f_\theta(y) = exp[(y\theta - b(\theta))/a(\phi) + c(y, \phi)],$$

where b, a, and c are arbitrary functions and ϕ is an arbitrary "scale" parameter. Moreover, $\mathrm{E}(Y) = b'(\theta) = \mu$ and $V(Y) = V(\mu)\phi$, where $V(\mu)$ denotes the variance function.

3.4.2.1 Socio-demographic model

Several authors have shown a list of the most common factors associated with AAD ([39, 53, 65, 81]). Those factors are mainly: being male and younger than 50 years old. The employment status, such as being unemployed, and the civil status, such as being single, also appear in some countries to be associated with AAD. A GAM is the appropriate statistical tool to find out what the factors associated with AAD are in the Portuguese population.

A possible definition, using model 3.2, called "GAM - Socio-demographic model", can be:

$$\begin{aligned} g(\mu_i) \quad = \quad & \beta_0 + s_1(d_i) + s_2(pd_i) + \beta_1 ge_i \\ & + s_3(a_i) + \beta_2 cs_i + s_4(e_i) + \beta_3 i_i \\ & + \beta_4 es_i, \end{aligned} \tag{3.3}$$

with $i = 1, \ldots, 2060$, $y_i = 1, 0$ (see subsection 3.2.8 for more details). In this case, $\theta_i = \ln(\mu_i/1 - \mu_i)$, the well known *logistic regression*. The response variable will take value 1 if the individual i suffers from AAD and will take value 0 if the individual i does not suffer from AAD. Very few cases were included in some categories, so, to prevent spurious conclusions, cases were trimmed,

- "Total mental disorders" is the total number of mental disorders per individual with d_i varying between 0 and 10 (trimmed),
- "Total parent's mental disorders" is the total number of mental disorders that each individual's parents suffer(ed) with pd_i varying between 0 and 5 (trimmed),
- "Gender" is a two-level categorical variable with $ge_i = 1$ if the subject is male and $ge_i = 2$ if the subject is female,
- "Age" is the age of the individual at the time of interview with a_i varying between 18 and 80 years of age,

- "Civil status" is a three-level categorical variable with:

$$cs_i = \begin{cases} 1, & \text{if } cs_i = \text{Married or Cohabiting,} \\ 2, & \text{if } cs_i = \text{Separated, Widowed, or Divorced,} \\ 3, & \text{if } cs_i = \text{Never Married,} \end{cases}$$

- "Education" is the number of educational years that each individual has with e_i varying between 0 to 17 years,

- "Income category" is a four-level categorical variable with:

$$i_i = \begin{cases} 1, & \text{if } i_i = \text{Low income,} \\ 2, & \text{if } i_i = \text{Average Low income,} \\ 3, & \text{if } i_i = \text{Average High income,} \\ 4, & \text{if } i_i = \text{High income,} \end{cases}$$

- "Employment status" is a three-level categorical variable with:

$$es_i = \begin{cases} 1, & \text{if } es_i = \text{Working,} \\ 2, & \text{if } es_i = \text{Retired,} \\ 3, & \text{if } es_i = \text{Other (incl. unemployed, disabled, students, etc.).} \end{cases}$$

For the four factor variables described above (Gender, Civil status, Income category and Employment status), identifiability constraints were imposed to prevent the creation of a not full column rank model. The first level (=1) of each the variables was removed from the model, implicitly treating the corresponding parameter as zero. The logistic regression coefficients ($\hat{\beta}_j$) were exponentiated to create Odds Ratio (OR) and 95% CI.

Some covariates have a smooth function, s_k, with $k = 1,\ldots,4$, instead of having a multiplicative parameter β because it is unknown if the continuous variables, "Total mental disorders", "Total parent's mental disorders", "Age", and "Education" enter the model linearly. The relationship between those variables and the AAD cases can thus be flexibly determined. Notice that smooth functions cannot be applied to non-continuous variables, and therefore the model involves parametric and non-parametric terms.

Figure 3.1 shows the smooth estimates for the continuous variables. From top-left to bottom-right the covariates represented are: Total mental disorders; Total parent's mental disorders; Actual age; and Educational years. The "rug plot", visible at the bottom of the "age" graph, is used to show the covariate values. The graphs allow visualizing the influence of each predictor variable in the response once the other predictors have been controlled for. The plots show the estimated effects as solid lines/curves. In the horizontal axis are the continuous variables on the scale of the predictor, on the vertical are the coefficients of the association. Between brackets are the

Estimated Degrees of Freedom (EDF) for each covariate. EDFs show the complexity of the non-linearity of the relationship, starting on one for linear relations and growing from there. The number of other total mental disorders positively impacts the existence of AAD. However, the impact is not linear. It is relatively constant between one and six other mental disorders and increases significantly thereafter, suggesting the existence of a potentially meaningful cut point in the covariate. The effect of mental disorders of parents of the individual is, in fact, a linear relationship, as proven not only visually but also by the number of EDFs, one. While those suffering from other mental disorders seems to suffer more from AAD, those whose parents suffer(ed) from mental disorders seems to suffer less from AAD. Younger people tend to suffer the most from AAD. The education factor has an impact, but only to differentiate between those who have no education at all from others who have at least an elementary education (four years of education), suggesting again the existence of a cut point in the covariate, demonstrating that a dichotomous categorization below and above that point has some validity to understand the relationship between the variables [14]. In fact, this result was expected due to the characteristics of the people with no education in Portugal, only a few people older than 65 years of age are in this situation (younger people suffer more AAD than do older people).

We can conclude from Table 3.5, that Portugal is much like other countries. In fact, younger males are those who suffer the most from AAD. Those who are "separated, widowed, or divorced" appear to suffer more than do married people, as in other countries. The employment status is not related to AAD in Portugal and those with an average income suffer less than those within the two extreme levels.

3.4.2.2 Co-morbidities model

It is clinically important to understand which mental disorders are most commonly present with AAD. To answer this question, a GLM was fitted having as covariates all other mental disorders.

A possible definition, using model 3.1, called "GLM - Co-morbidities model", can be:

$$\begin{aligned} g(\mu_i) \quad = \quad & \beta_0 + \beta_1 hyp_i + \beta_2 so_i + \beta_3 sp_i + \beta_4 mde \\ & + \beta_5 mddh_i + \beta_6 dys_i + \beta_7 mnd_i + \beta_8 pat_i \\ & + \beta_9 gad_i + \beta_10 sad_i + \beta_11 asa_i + \beta_12 pts_i \\ & + \beta_13 ald_i + \beta_14 cd_i + \beta_15 odd_i + \beta_16 odd_i \\ & + \beta_17 other_i, \end{aligned} \tag{3.4}$$

with $i = 1, \ldots, 2060$, $y_i = 1, 0$ (see subsection 3.2.8 for more details). Again in this case, $\theta_i = \ln(\mu_i / 1 - \mu_i)$, the *logistic regression* model. The response variable will take value 1 if the individual i suffers from AAD and will take value 0 if the individual i does not

Parameter	GAM	
	AOR	P-value
β_1 - Total number of mental disorders		0.00***
β_2 - Total number of parents mental disorders		0.09.
β_3 - 1. Male	1.00	
β_3 - 2. Female	0.05	0.00***
β_4 - Age		0.00***
β_5 - 1.Married or Cohabiting	1.00	
β_5 - 2. Divorced or Separated or Widowed	1.85	0.01**
β_5 - 3. Never Married	0.95	0.76
β_6 - Education		0.00***
β_7 - 1. Low income	1.00	
β_7 - 2. Average Low income	0.76	0.05.
β_7 - 3. Average High income	0.70	0.02*
β_7 - 4. High income	0.76	0.19
β_8 - 1. Working	1.00	
β_8 - 2. Retired	0.65	0.14
β_8 - 3. Other	0.75	0.09

Table 3.5: Socio-demographic model results - GAM (all covariates included). AOR - Adjusted OR to allow over dispersion. Significance codes: 0 "***"; 0.001 "**"; 0.01 "*"; 0.05 ".".

suffer from AAD. All covariates will also take value 1 if the individual i suffers from that specific mental disorder and will take value 0 if the individual i does not suffer from that mental disorder.

Due to the very low prevalence rate, some mental disorders were grouped in a class of "Others". These are: Mania, Panic Disorder, Obsessive-compulsive disorder, Agoraphobia with and without panic, Bulimia, Binge, Pre-menstrual Disorder, and Attention Deficit Disorder. A quasi-likelihood model was fitted and at a lower than 5% significance level the following disorders have an OR greater than 1: Alcohol Dependance (ald), Oppositional Defiant Disorder (odd), Hypomania (hyp), and Intermittent Explosive Disorder (ied). At the same level of significance, one disorder is not associated with AAD, with an OR lower than 1: Generalized Anxiety Disorder (gad). See Figure 3.2 for more details. Figure 3.2 shows the $\hat{\beta}_j$ values, and the standard errors associated with each of the mental disorders included in the GLM. The name attributed to each of the other mental disorders with AAD in Figure 3.2 and in model 3.4 follows the rule: d_disorder_our. The list of disorders is: hyp: Hypomania; so: social

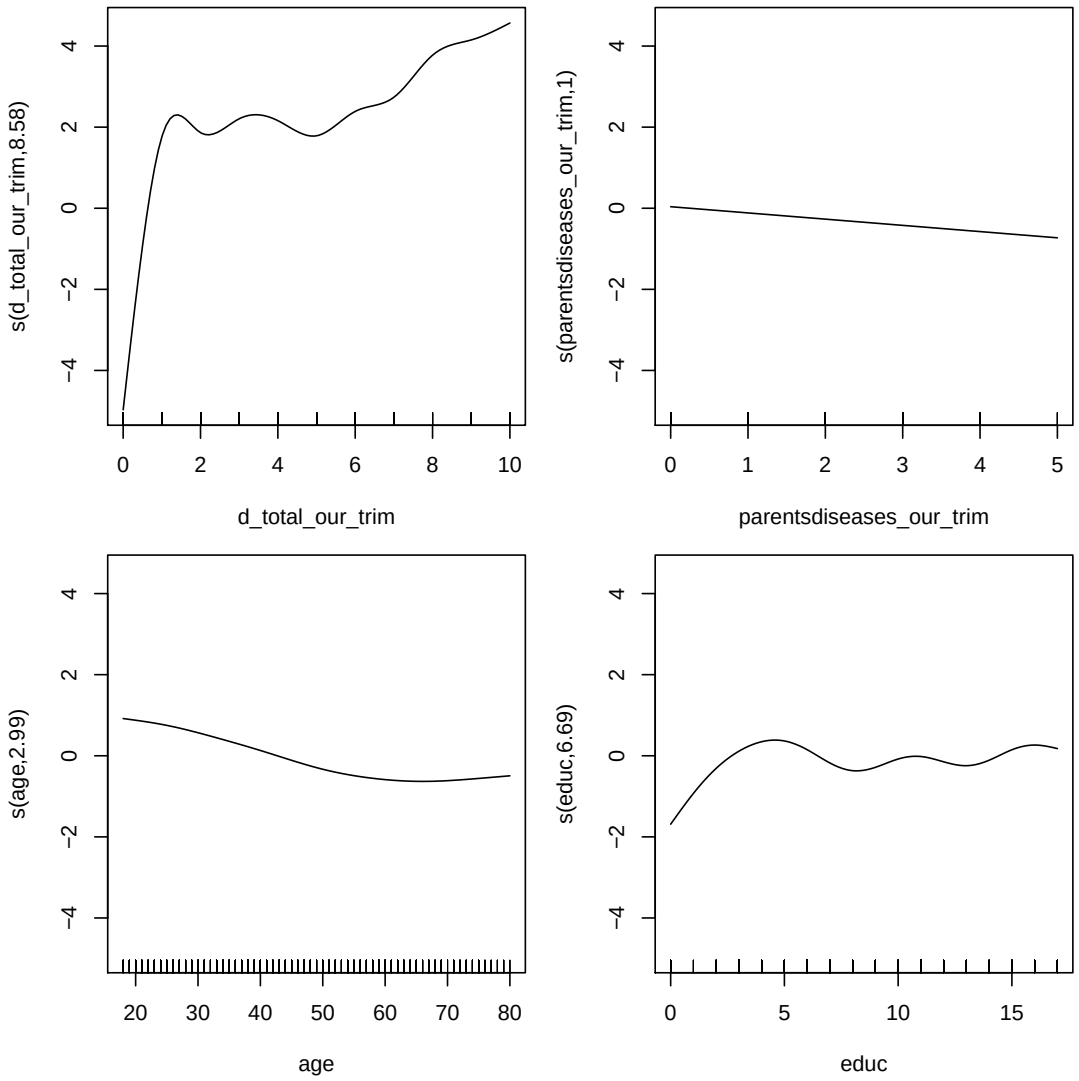

Figure 3.1: Smooth estimates for the selected continuous variables.

phobia, sp: specific phobia; mde: major depressive episode, mddh: major depressive disorder with hierarchical rules; dys: dysthymia; mnd: minor depressive episode; pat: panic attack; gad: generalized anxiety disorder; sad: separation anxiety disorder; asa: adult anxiety disorder; pts: post-traumatic disorder; ald: alcohol dependence disorder; cd: conduct disorder; odd: oppositional-defiant disorder; idd: intermittent-defiant disorder.

3.5 Alcohol Abuse Disorder distribution across Portugal

In this Section models presented in Chapter 2 are applied to the Portuguese AAD data collected.

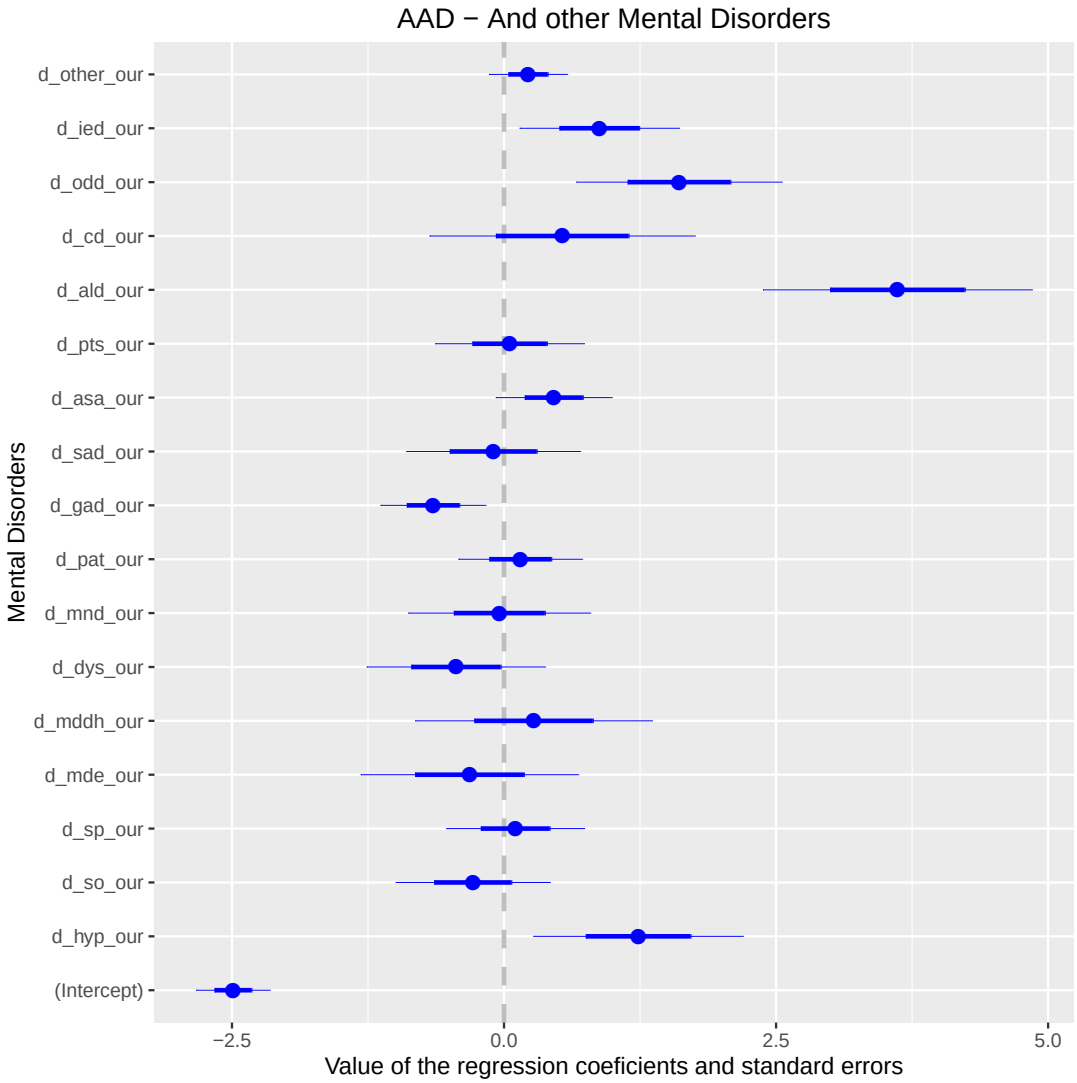

Figure 3.2: $\hat{\beta}_j$ values, and the standard errors associated with each of the mental disorders included in the GLM.

3.5.1 Standardized morbidity ratio

The study region is mainland Portugal partitioned into 28 units called Nomenclatura Comum das Unidades Territoriais Estatísticas (NUTS3)[2], corresponding to the 3rd level territorial units aggregation. There are 30 NUTS3 in Portugal, from which 28 are in mainland Portugal and 2 are in the Islands. The *response variable* is the number of lifetime AAD cases per NUTS3. Differences in the size and demographic structure of the population living in each NUTS3 are accounted for by computing the expected number of AAD cases using indirect internal standardization. Age-specific rates for the disease at each NUTS3 are not available and therefore the indirect method is used,

[2]In Portuguese as defined by Eurostat, the European statistical organization.

by applying the age-specific disease rate for the global population to the NUTS3 age-specific population, provided by INE for the year of 2008. As this standardization is done using the AAD age-specific disease rate for the global population, as it was collected by the survey itself, the standardization is internal.

Figure 3.3 shows the raw SMR values for the 28 NUTS3.

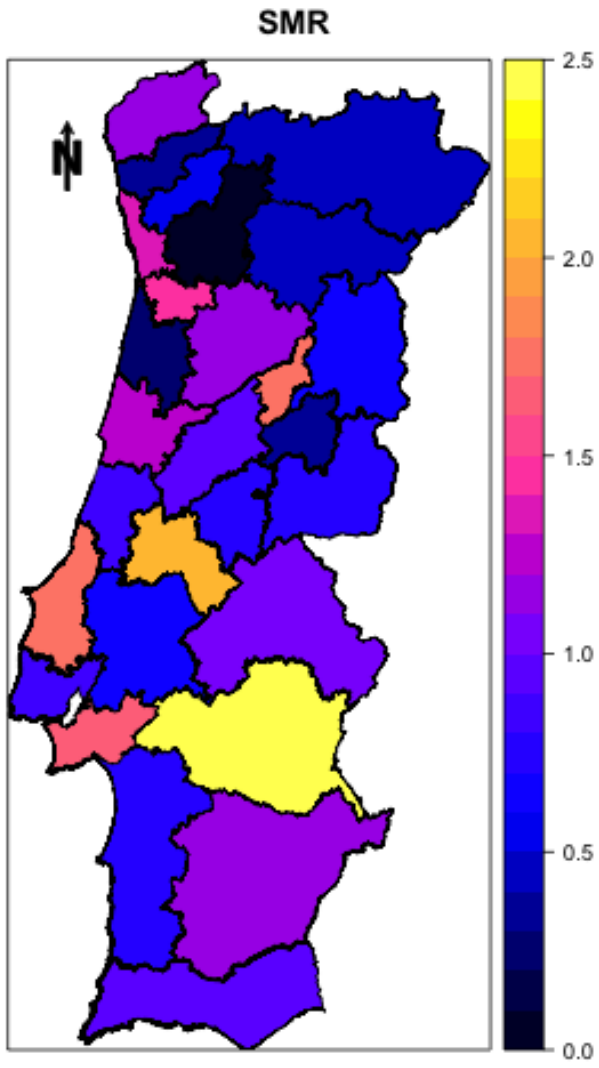

Figure 3.3: AAD Raw SMR per NUTS3. The four regions, which had originally missing values, are shown already with the imputed mean values resulting from the GLM (see Subsection 3.5.2).

Our illustrative example also considers two ecological covariates that, as seen before, are widely known as being associated with AAD, which are (a) proportion of population aged 18 to 34, (b) proportion of males. These data are only available per NUTS3, for the year of 2011, as provided by the latest census conducted in Portugal, which we find to be temporally misaligned with WMHSI data used in this work. However as population age and gender structures do not significantly change in 3 years, no

corrective measures have been implemented.

3.5.2 Disease mapping

The number of lifetime AAD cases vary between 2 679 (16A - Cova da Beira) and 136 789 (171 - Grande Lisboa). There are four NUTS3 (164 - Pinhal Interior Norte, 166 - Pinhal Interior Sul, 169 - Beira Interior Sul and 181 - Alentejo Litoral) where no cases were identified. The national nature of the survey sampling design creates situations where very small or even zero samples at the NUTS3 level occur. In this situation it might happen that no cases are estimated, which does not mean that no disease diagnoses exist. Therefore, these areas are treated as having missing values and not as having a null number of cases. The first level of the Bayesian hierarchical model, as seen in (2.1), involves complex calculations, very difficult to run on such numbers, therefore numbers of cases per 100 inhabitants, as well as expected number of cases per 100 inhabitants are used (this change does not eliminate the need of using the expected number of cases because only the size of the population is accounted for, not the structure).

The R software (version 3.1.1), with the package **CARBayes** [48] is used to fit the hierarchical models. The main advantages of this package are: (1) the spatial adjacency information is easy to specify as a binary neighborhood matrix; (2) given the neighborhood matrix the models can be implemented by a single function call in R; (3) maps with the disease risk estimates can easily be produced. The package has predefined the following models that will be used: BYM, LLB and LCAR. By running the same model on R and on the BUGS software [57]) the package's author shows that there is good agreement between the two sets of point estimates, as we confirm in the present work. One disadvantage of the package is that it cannot handle missing values at the response variable level. To overcome this, a GLM, Poisson (quasi-likelihood) model [61]), is fitted using as response variable the number of lifetime observed cases per NUTS3 and as covariates the ecological variables defined before, namely the proportion of men and the proportion of population aged 18 to 34. The mean estimated number of lifetime observed cases achieved for the four areas with missing data are incorporated in the response variable vector $\mathbf{Y}$. This methodology is debatable and more work would have need to be done, in order to evaluate all possible consequences of this approach, if in the meanwhile authors of the package would have not changed it (see following chapters).

The MBYM model is fitted using the OpenBUGS software [57]). Even though the Bayesian methodology could handle the missing values, for comparison purposes the missing values are also replaced by the mean estimated values.

As mentioned in Subsection 2.4.3, the authors of LCAR propose that, for the elicitation of $\tilde{\boldsymbol{W}}$, data having a similar spatial structure as the response variable should be used. In their case, the prior elicitation was based on response variable data from

previous years. Our decision was to use the number of cases of four other related mental disorders, chosen as per model 3.4 and therefore using the following disorders as covariates: Alcohol Dependence, Oppositional Defiant Disorder, Hypomania, and Intermittent Explosive Disorder. In the cases where values are missing the procedure followed is the one defined before, using as covariates the remaining disorders. For example, for alcohol dependence disorder as response variable, the covariates are: alcohol abuse disorder, oppositional defiant disorder, hypomania and intermittent explosive disorder. The mean estimated number of cases are imputed in the response variable vectors. There are two reasons to use a different approach in the present case. First, in Portugal, data on AAD from previous surveys is not available. Second, this work is on DM and not on ESR, therefore the decision is to use data from related mental disorders.

3.5.2.1 Hyperpriors

Table 3.6 shows the prior distributions implemented in the four models. In the LCAR model, on top of the already mentioned information for the $\tilde{\boldsymbol{W}}$ matrix, the parameter ϵ is set to 0.001.

Model	Parameter	Prior Distribution	Mean/Shape	Variance/Scale
BYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	$\boldsymbol{\mu}$	Gaussian	0	1000
	σ_θ^2 and σ_ψ^2	Inverse-Gamma	0.001	0.001
MBYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	100000
	$\boldsymbol{\mu}$	Flat		
	σ^2	Inverse-Gamma	0.001	0.001
	λ	Uniform [0,1)	0.5	0.5
LLB	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Inverse-Gamma	0.001	0.001
	ρ	Uniform [0,1)	0.5	0.5
LCAR	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Uniform [0,1000)	500	500

Table 3.6: Prior distributions for the models.

3.5.2.2 Inference

Posterior inference for all models is based on McMC simulation, using a combination of Gibbs sampling and Metropolis-Hastings algorithms. Posterior inference is based on 8 000 McMC samples, which are obtained by running one chain for 100 000 samples, by which convergence is assumed to have occurred. We ignore the first 20 000 samples as burn-in, and use the remaining 80 000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to reduce the autocorrelation).

Pilot runs are carried out to establish appropriate burn-in using the Geweke's diagnostic [29]). Convergence is assessed by visually monitoring the trace and the posterior density plot for each of the parameters. See figure 3.4 for an example of the convergence plots.

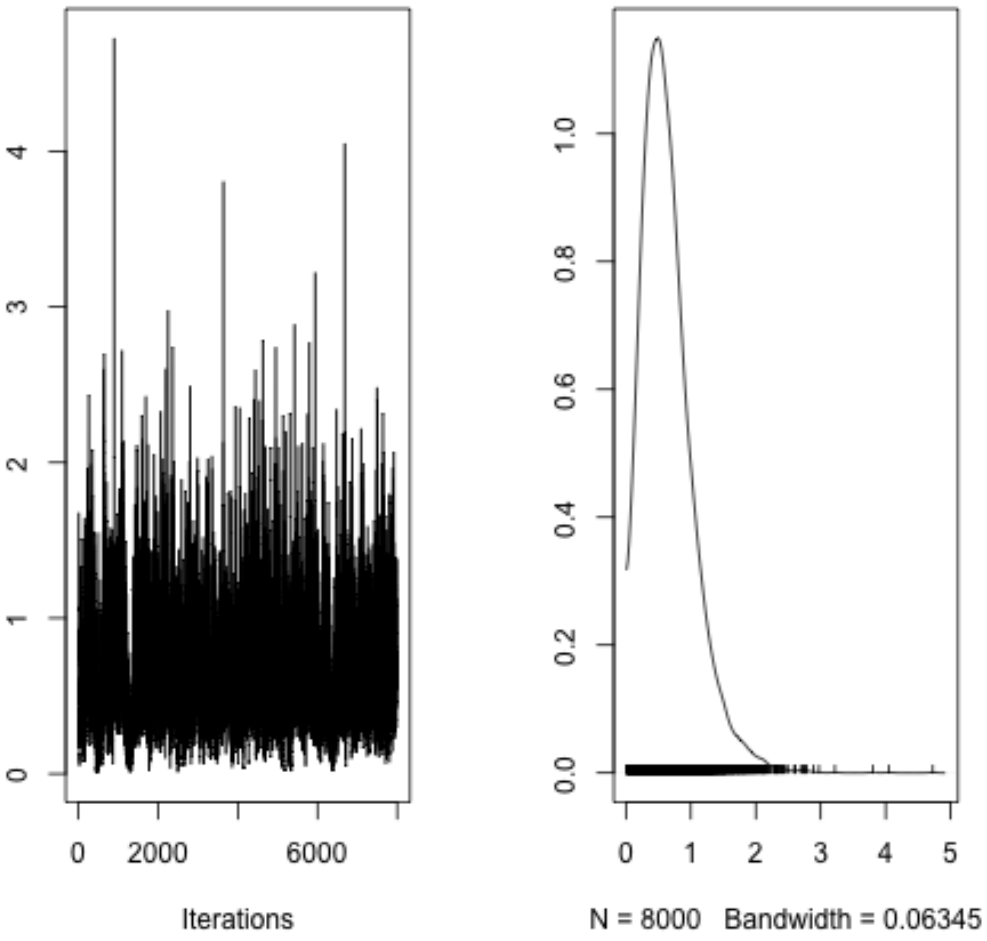

Figure 3.4: Trace and density plot for one of the parameters of the model.

3.5.2.3 Results

Each model is assessed by the resulting Deviance Information Criterion (DIC) [79], where a smaller value represents a better fitting model. Table 3.7 shows the results of the four models.

Table 3.7 shows that, according to DIC, the MBYM model exhibits the best fit. BYM model is the second best. Following MacNab [59], $\lambda = 1$ represents spatial/local smoothing and $\lambda = 0$ represents non-spatial/local smoothing, based on the disease

	BYM	MBYM	LLB	LCAR
DIC	155.3	145.0	159.2	158.0
p_D	14.3	5.8	18.5	19.5

Table 3.7: DIC results, which include the effective number of parameters in the model (p_D).

mapping data at hand. In the MBYM model the posterior mean value of $\lambda = 0.58$, shows that the data has an higher spatially structured variance than unstructured variance. As already proved by Lee [47], the BYM model shows more robust results in the presence of strong spatial correlation structures, as it seems to be the case here.

Figure 3.5 shows the posterior median SMR values for the 28 NUTS3, produced by the MBYM model. Table 3.8 shows summary measures of the marginal posterior of the parameters of interest obtained by the MBYM model.

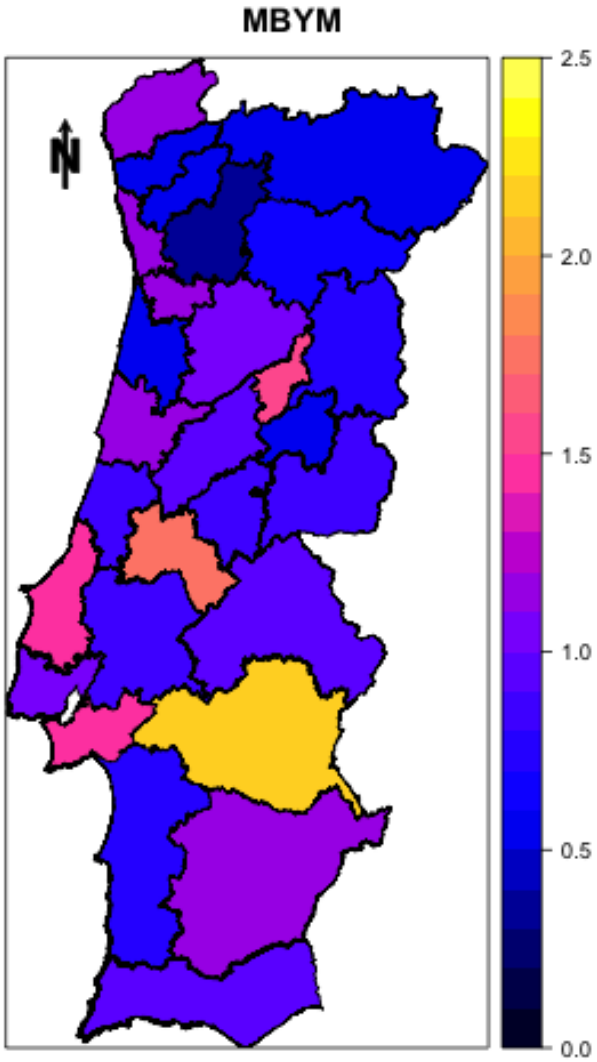

Figure 3.5: MBYM AAD posterior median SMRs per NUTS3.

Parameter	Prior distribution	Prior mean	Prior std	MCMC Posterior mean (std)	2.5%	Median	97.5%
β_0	Flat	0		-0.11 (0.10)	-0.32	-0.11	0.08
β_1	Gau (0, 100000)	0	100000	-0.23 (0.14)	-0.52	-0.22	0.06
β_2	Gau (0, 100000)	0	100000	-0.8 (0.13)	-0.34	-0.07	0.18
λ	U [0,1)	0.5	0.5	0.58 (0.25)	0.07	0.61	0.97
σ^2	IG (0.001, 0.001)	1	10	0.61 (0.17)	0.35	0.59	1

Table 3.8: MBYM model parameters summary.

Figure 3.6 displays histograms of the (a) raw SMR and the (b to d) smooth posterior median SMR values for the 28 NUTS3, produced by the models. The concentration around the interval [0.5, 1.5] on the latter can clearly be seen. Mapping the raw SMRs gives a misleading picture of the risk pattern, whereas any of the four models (plus LLB, which is not presented, but shows the same overall results) give posterior median relative risks less dispersed. This ability of the Bayesian models to "clean" adequately the SMRs from the false patterns created by the Poisson noise had been already referred by Richardson et al. [73].

3.5.2.4 Conclusions

In this Section the models presented in Chapter 2 are used to estimate the disease risk of AAD at the NUTS3 level, in Portugal.

In terms of DIC, the MBYM model achieves the best results. In the present case its superior performance is likely to result from the BYM (and MBYM) model ability of achieving the best results in cases when the spatial correlation structure is strong, as seems to be this case. The LLB model has consistently shown good results across a variety of cases but in this study, in terms of DIC, it proves to be the most poorly performing. While other authors show that the LLB model is the one achieving the best results [47, 59], our study shows otherwise. The performance of each model will depend on the type of data at hand, and none can be defined as the 'gold standard' over others.

The LCAR model is the only one that does not have a single global level of smoothness and therefore any existing discontinuities in the risk pattern can only be concluded from this model. There are 122 neighborhoods (or connections) between the 28 NUTS3. When applying the LCAR model, the 95% credibility interval of the number of removed connections is [2, 54]. This fact provides evidence that there is information in the data to estimate the number of connections to be removed. Results confirm the known deep cultural roots in the country on the differences between the coast- and the country-side NUTS3. This is the case of Península de Setúbal and Algarve, two coast-side NUTS3 sharing physical borders with the country-side NUTS3 Alentejo, which are no longer present when data is used to estimate connections.

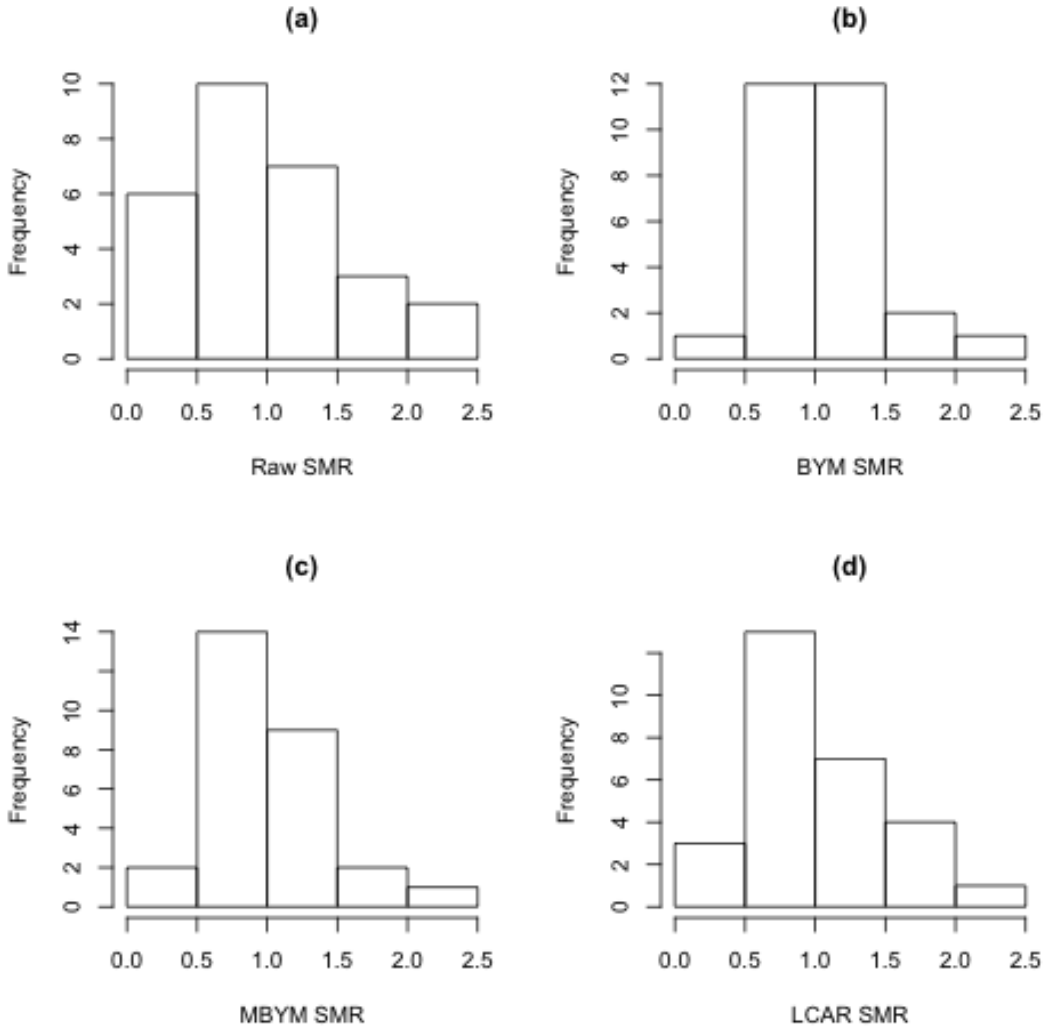

Figure 3.6: Histograms of the (a) raw SMRs and posterior medians of the (b,c,d) SMRs, for all areas derived by each of the three models, (b) BYM, (c) MBYM and (d) LCAR.

As mentioned in Subsection 2.2.2 the goal of DM is not the estimation of associations between covariates and the disease cases, but is to estimate the pattern of disease risk over a geographical region. Nevertheless, due to the fact that the two coefficients (β_1 and β_2) did not show to be significantly different from zero (contrary to expectations mentioned in Section 2.3), one must remember that this is an ecological study design, and the results must not be interpreted in terms of individual level cause and effect. One possible explanation is ecological bias as the prevalence rate of AAD is higher in younger men. Another possible explanation is that both the random and the covariate effects are confounded, because both are globally smooth in the MBYM model.

This study has some particularities when compared with the majority of the published applications:

a. the data use emerged from a survey, which was not plan to have local (at NUTS3 level) samples with the proper size to allow designed-based estimation, and therefore presents some missing values. To overcome this a frequentist model is used;

b. the complex computations of the first level of the hierarchical Bayesian models do not allow the direct use of the survey estimates. To overcome this the number of lifetime cases of AAD per 100 inhabitants is used;

c. the LCAR model is used as a DM and not as a ESR, and therefore the type of data used for the elicitation of the $\tilde{\mathbf{W}}$ matrix is not previous periods data for the same disease but data from correlated disorders.

The epidemiological study presented in this work shows substantial evidence of some *"hot spots"* in the Center and South of the country allowing the authorities to focus interventions on these "excess risk" areas.

Chapter 4

A Gaussian random field model for similarity-based smoothing

4.1 Introduction

The contents of this chapter is based on the paper: **A Guassian random field model for similarity-based smoothing in Bayesian Disease mapping [6].**

For a long time, DM models have had a basic fundamental understanding: "...the values for a pair of contiguous zones would be generally much more alike than for two arbitrary zones,..." [9]. This concept has become so important that we can find the following definition of DM: [84] (our underlining)

"In the statistical literature, "disease mapping" refers to a collection of methods extending small area estimation to directly utilize the spatial setting and assumed positive spatial correlation between observations, essentially borrowing more information from neighboring areas than from areas far away and smoothing local rates toward local, neighboring values."

Results from the models presented in Section 3.5 seem to confirm that concept as the mean value of $\lambda = 0.58$ indicates a spatial/local smoothing variance higher than unstructured variance. Nevertheless, intuitively and just by looking at Figure 4.1, the spatial positive correlation seems much less marked in Portugal than in Scotland. On top of that, in the words of Best et al. [11] (pag. 145) the BYM model "...shows a tendency to over-smooth the data and to exhibit spurious geographical patterns in the area-specific risk estimates when no underlying excess risk is present" and this may lead to (pag. 144) "the danger of over-interpreting apparent spatial variation in disease rates in the absence of a clear *a priori* hypothesis of spatial variation in the underlying cause". Therefore, we returned to the previous step, and looked for a

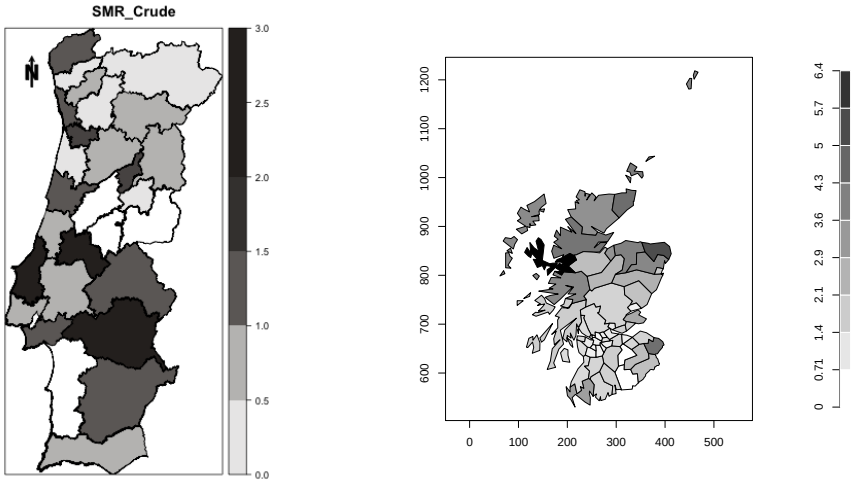

Figure 4.1: Crude SMR for (left hand map) AAD in Portugal as collected by WMHSI and (right hand map) lip cancer in Scotland.

formal confirmation of the spatial positive correlation in the AAD Portuguese data.

We could not find in the Portuguese AAD data that positive spatial correlation (see Subsection 4.6.1) but we still needed to borrow information from other areas, so, our thinking process moved us away from the above definition to find a possible solution for those datasets where the "traditional" way of thinking was not helpful.

The borrowing information mechanism is achieved by using the conditionally autoregressive CAR random effects distribution defined by Besag et al. [9]. This distribution uses a neighbourhood matrix built in a way that weight is given to the contiguous areas and no weight to the remaining ones. As this structure is very well established we wanted to take advantage of that by changing the matrix.

In Section 4.2, the hierarchical Bayesian model BYM and the CAR prior, are briefly presented again, and the actual neighbourhood matrices definitions are detailed. In Section 4.3 the new approach for the weight matrix, the similarity-based GRF matrix, is explained. A simulation study is presented in Section 4.4 to demonstrate the performance of the proposed model. In Section 4.5 some background information about alcohol abuse determinant factors is provided. The two illustrating case studies are presented in Section 4.6. Finally, Section 4.7 presents a summary discussion.

4.2 BYM, CAR and *Neighbours* definition

In this section we briefly describe the BYM model introduced in Chapter 2 together with the CAR prior. A general formulation of the likelihood of a Bayesian hierarchical model is given by

$$Y_i|E_i,R_i \sim Poisson(E_iR_i) \text{ for } i=1,...,n,$$
$$\ln(R_i) = \mu + \boldsymbol{\beta}\mathbf{x_i^T} + \phi_i. \tag{4.1}$$

The study region is partitioned into n small areas labelled $i = 1,\ldots,n$. Conditioning on the relative risk R_i the number of disease counts Y_i is assumed to be Poisson distributed with mean E_iR_i, where E_i is the expected number of cases in each area i computed using some kind of standardization based on the size and demographic structure of the population living in each area.

The log risks are represented by an intercept term denoted by μ, the vector of p covariates denoted by $\mathbf{x}_i^T = (x_{1i},\ldots,x_{pi})$ multiplied by the corresponding vector of regression parameters $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$, and a random effect ϕ_i, serving to quantify the effects of unmeasured covariates or confounders and also to account for the residual variation unexplained by the included covariates $\mathbf{x}_i^T$.

The BYM model defines $\boldsymbol{\phi}$ in model (4.1) by

$$\phi_i = \theta_i + \psi_i,$$

$$\theta_i|\sigma_\theta^2 \sim N(0,\sigma_\theta^2),$$

$$\psi_i = (\psi_1,\ldots,\psi_n)|\mathbf{W},\sigma_\psi^2 \sim ICAR(\mathbf{W},\sigma_\psi^2), \tag{4.2}$$

in which, θ_i represents a randomly varying component and assumes an independent and identically distributed normal prior, while ψ_i represents a spatially varying component and assumes an ICAR prior. Instead of a specification of a single multivariate distribution $f(\boldsymbol{\phi})$, CAR models are specified by a set of univariate full conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1,\ldots,\phi_{i-1},\phi_{i+1},\ldots,\phi_n)$. More detailed specifications can be found elsewhere [9].

Different strengths of spatial correlation can be represented by varying the relative sizes of the two components $(\boldsymbol{\theta},\boldsymbol{\psi})$. This flexibility is also a disadvantage, as each data point is represented by two random effects while only their sum $(\theta_i + \psi_i)$ is identifiable [59].

W is the neighbourhood matrix. The most common types of neighbourhood matrices used are two: (a) the adjacency-based GMRF matrix, defined as

$$w_{ij} = \begin{cases} 1, & \text{if } j \sim i \\ 0, & \text{otherwise,} \end{cases}$$

where $j \sim i$ represents contiguous areas, and therefore j and i are considered *neighbours* (elements w_{ii} are equal to zero and $w_{ij} = w_{ji}$ [9]), and will be named the **W**-based GMRF matrix; and the (b) distance-based GMRF matrix, defined by Best et al. [11], as a binary $n \times n$ matrix, with elements $d_{ij} = e^{-k_{ij}/\delta}$, for k_{ij} = distance (in kilometres) between the geographic centroids of area i and j, and δ is chosen to give a relative weight of 1% ($d_{ij} = 0.01$) to an area j whose centroid is equal to the mean inter-district distance for the study area, from area i. Elements d_{ii} are equal to one and $d_{ij} = d_{ji}$. Herein after this matrix will be named the **D**-based GMRF matrix.

4.3 A similarity-based Gaussian random field model

The GRF model proposed herein no longer retains the Markovian properties as those based on the neighbourhood weights. Instead of using spatial distance or spatial adjacency, a measure reflecting similarity between areas is introduced. This requires a deep knowledge of the disease data at hand, and therefore cannot be governed by convenience and/or convention, as has been the case until now [32]. Data used should come from: a) a disease determinant factor or a combination of factors, b) a source external to the survey that collected the disease data. The main objective of the proposed model is the provision for borrowing strength between areas with similar disease determinant factors.

Firstly, regions exhibiting the "same"/close level of risk in a determinant factor will be regions with the "same"/close level of risk of the disease. Secondly, if disease data need to be *strengthened*, using disease determinant factor information collected by the same survey might inflate or not remediate possible *weaknesses* of the disease data. Therefore, an external source for the disease determinant factor is critical.

The rationale of our approach is the following: in cases of diseases with no environmental determinant factors, use of a positive spatial correlation based on physical distance or adjacency, in the GMRF model, may not be the best way to reflect similarity between areas. By using the GRF model reflecting *how similar* each area is to one another, in terms of a disease determinant factor that was collected by an external source, the disease risk distribution can be better assessed.

A distance-based matrix seems to be performing generally better than an adjacency-based matrix [11, 24]. Therefore, based on a matrix definition proposed by Best et al. [11], the new matrix, further denoted as **S**-based GRF matrix, with elements s_{ij} for each region i, has the following structure:

$$s_{ij} = \begin{cases} e^{-p_{ij}/\delta}, & \text{if } j \neq i \\ \frac{1}{n-1}\sum s_{(-i)}, & \text{otherwise,} \end{cases}$$

where p_{ij} is the absolute gap between region i and region j,

$$p_{ij} = |p_i - p_j|, \tag{4.3}$$

in terms of the disease determinant factor, and δ is equal to a value that gives a relative weight of 1% ($s_{ij} = 0.01$) to an area i whose difference from an area j is the mean inter-region difference for the country. Elements s_{ii} need a specific definition, otherwise their value would be the one contributing the most to the prior, as $e^0 = 1$ and all other s_{ij} elements have values between 0 and 1. Therefore, p_{ii} values are equal to the average value of all elements except the ith area value.

4.4 Simulation study

This section presents a simulation study that compares the performance of a BYM model a) with an adjacency-based GMRF **W** matrix and b) with a similarity-based GRF **S** matrix, when no positive spatial correlation is displayed by the disease data. The main goal of the simulation study is to assess the performance of a GRF model by using disease determinant factor data in two different ways, a) as a covariate or b) defining the weight matrix.

4.4.1 Study design

Simulated disease data are generated for 100 hypothetical areas using a regular 10×10 grid. The disease counts are generated from the BYM model (see Section 4.2 for details). Two independent normally distributed, $N \sim (0, 1)$, covariates were considered with a regression parameter of $\beta_1 = \beta_2 = 0.1$, while the intercept term is fixed at $\mu = -0.2$. The expected number of cases, **E**, are fixed at 40. To create the matrix mentioned in Section 4.3, disease determinant factor data (dd) were simulated through the realization of a 100×1 vector following a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = 0$ and covariance matrix covering the following scenarios:

1. Independence: Covariance matrix with main diagonal variances generated from a $Gamma \sim (1, 1)$ distribution.

2. Moderate correlation: Covariance matrix with main diagonal variances generated from a $Gamma \sim (1, 1)$ distribution and 0.5 constant correlation coefficient.

3. Strong correlation: Covariance matrix with main diagonal variances generated from a $Gamma \sim (1, 1)$ distribution and 0.8 constant correlation coefficient.

The adjacency-based GMRF **W** matrix is a first order contiguity Rook matrix, using only common boundaries to define the *neighbours*, with no vertices included [8]. For example, area 1 has areas 2 and 11 as *neighbours*.

The so-called structured random effect, $\boldsymbol{\psi}$, was generated by the process: $0.9 \times dd + \epsilon$, where ϵ follows a Gaussian distribution $N \sim (0, 0.05)$. To avoid any spatial structure (in the traditional sense) the 100 values were uniformly assigned to the 100 areas. The so-called unstructured random effect, $\boldsymbol{\theta}$, was generated through a vector of 100 independent Gaussian variables $N \sim (0, 0.2)$.

Five hundred sets of disease counts were generated under each of the three scenarios, and the BYM model, as defined in model 4.2 with the two different matrices, was applied in each case. Each simulated data set is generated from a different realization of the random effects as proposed by Lee [47], because it prevents the results from being affected by the particular set of random effects drawn. The relative performance of the two models is assessed by bias and Root-mean-square error (RMSE) for the

Metric	Scenario	Model		
		Adjacency	Distance	Similarity
% Bias	1	-0.77	-0.77	-0.91
	2	-0.65	-0.62	-0.74
	3	-0.69	-0.72	-0.82
RMSE	1	3.52	3.55	3.14
	2	3.63	3.72	3.00
	3	3.57	3.55	2.98
%Coverage probability	1	94.16	92.94	96.58
	2	95.64	94.81	97.58
	3	94.73	93.37	97.36

Table 4.1: Summary of the simulation study results for the estimated values of the disease risks E_iR_i. The bias and the coverage probabilities are presented as a percentage of the true values, while the RMSE is presented as the absolute difference to the true values. The coverage probabilities were calculated based on the 95% credible interval. One hundred simulations were carried out for the distance-based GMRF model.

estimated values, E_iR_i, which are presented as a percentage and absolute difference, respectively, of theirs true values. In addition, the coverage probabilities of the 95% credible interval for the E_iR_i values are again presented on the percentage scale. First the average coverage probability rate is calculated for each area for the 500 hundred simulations and secondly the summary statistics across the 100 areas are presented.

The BYM model with the similarity-based GRF **S** matrix includes the two covariates, while the model with the adjacency-based GMRF **W** matrix (and the distance-based GMRF **D** matrix) includes three covariates, the previous two plus the disease determinant factor data.

Inference, priors, and hyperpriors used for running the models are the ones used in the case studies (see Section 4.6).

4.4.2 Results

Results from all metrics and models are shown in Table 4.1. Overall, all scenarios (over the 500 data sets) produce close to unbiased estimates of the risks, with similarity-based GRF **S** model showing a slightly higher value, and values ranging between -0.74% and -0.91%. The similarity-based GRF **S** model performed the best in terms of RMSE. In the presence of strong correlation between the disease cases and the disease determinant factor data, the RMSE reaches its lowest value, and progresses inversely to the correlation coefficient. Table 4.2 shows the summary statistics for the performance, in terms of the average coverage probabilities, in detail. Overall, the coverage probabilities for the similarity-based GRF **S** model are above those of the adjacency-based GMRF **W** model.

Metric	Scenario	Model		
		Adjacency	Distance	Similarity
% Minimum	1	90.8	85.0	94.4
	2	92.2	84.0	96.0
	3	92.0	88.0	95.0
% First Quartile	1	93.4	91.0	96.0
	2	95.0	93.0	97.2
	3	93.8	91.8	97.0
% Median	1	94.2	94.0	96.6
	2	95.6	95.0	97.7
	3	94.6	94.0	97.4
% Third Quartile	1	95.1	95.0	97.2
	2	96.6	97.0	98.0
	3	95.6	95.0	97.8
% Maximum	1	97.8	99.0	98.0
	2	98.6	100.0	98.8
	3	97.4	98.0	99.2

Table 4.2: Summary of the simulation study results for the estimated values of the disease risks E_iR_i on the coverage probabilities for the 100 areas across the five hundred simulations. One hundred simulations were carried out for the distance-based GMRF model.

The first finding of this study is that in cases in which the disease data do not show a positive spatial correlation, it is more efficient to use the data from the disease determinant factor to build the similarity-based GRF matrix than to use it as a covariate. Efficiency gains result from the fact that the matrix helps to model the variability attributable to the effects of possibly omitted covariates that may not be spatially structured. The second finding of this study is that the capabilities of the GMRF model, with an adjacency-based matrix in the case of positively spatially correlated data, can be extended to the case of not positively spatially correlated data by changing the matrix base.

Table 4.1 and Table 4.2 show in the "Model" middle column the results of 100 simulations using the so-called distance-based GMRF **D** matrix. The results obtained with the BYM model with this matrix are very close to those obtained from the same model with the adjacency-based GMRF **W** matrix. Therefore the similarity-based GRF **S** matrix model shows the same advantages and drawbacks versus the two most used matrices types, the adjacency-based GMRF **W** and the distance-based GMRF **D**.

4.4.3 Results under different prevalence scenarios

To evaluate the performance of the S-based GRF model when applied to disease data with different prevalences, an extra simulation was conducted. This extra simulation

Metric	Scenario		
	Low $[5, 15]$	Medium $[35, 50]$	High $[50, 65]$
% RMSE	0.07	0.21	0.23
% Bias	-5.17	-2.95	-2.69
% Coverage probabilities (median over all areas)	97.00	88.00	92.50

Table 4.3: Summary of the simulation study (for different prevalence scenarios) results for the estimated values of the disease risks $E_i R_i$. The RMSE and bias are presented as percentages of the true values. The coverage probabilities were calculated based on the 95% credible intervals.

follows the same definition as before (except the so-called unstructured random effect, θ, which is now generated through a vector of 100 independent Gaussian variables $N \sim (0, 0.08)$). The expected cases are uniform random draws from the following three intervals: [5,15], [35,50], and [50,65]; One hundred models were run for each of the prevalence intervals.

Table 4.3 shows the results of the three scenarios. All values are shown as a percentage of the true values. As expected the methodology is equally efficient for the three prevalence types, although the bias decreases as the number of prevalent cases increases. The methodology seems to produce acceptable results across all the spectrum of prevalence scenarios.

4.5 A motivating example

While Degenhardt et al. [21] and our results (see Chapter 3) show that males are more likely than females and younger adults are more likely than older adults to have used all drug types (including alcohol), defining the disorder determinant factors as age and gender (both intrinsic determinant factors), the American Psychiatry Association [2] mentions that genetic factors explain only part of the risk, with a significant part of the risk for alcohol dependence[1] coming from environmental or interpersonal factors that might include:

a. cultural attitudes toward drinking and drunkenness,

b. the availability of alcohol (including price),

c. expectations of the effects of alcohol on mood and behaviour,

d. acquired personal experiences with alcohol,

[1] DSM-V, the most recent edition of the manual, combines alcohol abuse and alcohol dependence into a single disorder, measured on a continuum from mild to severe.

f. and stress.

Cultural attitudes toward drinking and drunkenness could be considered an extrinsic factor, in a multi-cultural and multi-ethnic country, but would not be so considered in Portugal, an almost mono-ethnic state [54]. Regarding alcohol availability, the legal framework applies in the same way throughout all of the mainland Portugal. It may thus be considered as extrinsic, but only in countries with a decentralized legal structure, not the case in Portugal [4]. The remaining three factors mentioned are intrinsic. Connor et al. [18] mentions that some studies have estimated that 50-70% of the risk of alcohol use disorders is attributable to additive genetic factors.

Balsa et al. [4] is the latest published study on the population alcohol consumption in Portugal and refers to the period of 2007. We used alcohol use lagged by one year relative to alcohol abuse, as it ensures that alcohol use occurred before disorder onset. The percentage of the population in each NUTS 3 that regularly consumes alcohol is used in the motivating example as explained in subsection 4.6.3.

4.6 Case studies

4.6.1 Assessing spatial structure

To assess the presence of residual spatial autocorrelation, two overdispersed Poisson GAM models are fitted to the disease count data. The first one includes covariates and the second one does not. The first model is used to measure if any spatial correlation has not been accounted for by the available covariate information [47]. The second model is used to measure if the disease count data have a positive spatial correlation. The model is given by

$$Y_i|E_i, R_i \sim Poisson(E_i R_i) \text{ for } i = 1, ..., n,$$

$$ln(R_i) = \mathbf{S}(x_i^t) \tag{4.4}$$

or

$$ln(R_i) = \beta_0, \tag{4.5}$$

in the second case. This model assumes that disease counts are independent conditional on the available covariates. This is the same model as show in Section 4.2, model 4.1 without the random effects and the regression parameters replaced by smooth functions [91] in the model 4.4. A permutation test based on Moran's I statistic [63] (see subsection 2.5.3) using 10 000 random permutations was conducted in the raw residuals of the models.

4.6.1.1 Portuguese alcohol abuse data

Model with Covariates: Study region is partitioned into $i = 1, ..., n, (n = 28)$ NUTS3 in Portugal. Total number of alcohol abuse cases in area i is denoted by y_i, while the e_i is

the expected number of cases in the same area. A vector of p covariates is denoted by $\mathbf{x}_i^T = (x_{1i}, ..., x_{pi})$ (including a column of 1s for the intercept term) and is multiplied by a vector of smooth functions $\mathbf{S} = (S_1, ..., S_p)$. Smooth functions used in the model are natural cubic splines [91]. Included covariates are the proportion of population aged 18 to 34, the proportion of males, and the number of regular users, while R_i denotes the risk of disease in area unit i.

The number of regular users and the proportion of population aged 18 to 34 revealed substantial relationships with alcohol abuse disorder, and were thus retained in the model (number of degrees of freedom were, respectively, 2 and 3).

Statistically insignificant spatial autocorrelation was observed, with the Moran's I statistic equal to -0.2309 and a corresponding p-value for the null hypothesis of no spatial correlation of 0.97.

Model without Covariates: Again, statistically insignificant spatial autocorrelation was observed, with the Moran's I statistic equal to 0.0758 and a corresponding p-value for the null hypothesis of no spatial correlation of 0.16.

4.6.1.2 Scotland lip cancer data

The study region is partitioned into $i = 1, \ldots, n, (n = 56)$ districts in Scotland. Total number of lip cancer cases in area i is denoted by y_i, while the e_i is the expected number of cases per district calculated accounting for the different age distributions in each district. Included covariate is the percentage of the workforce in each district employed in agriculture, fishing, and forestry (AFF), modelled linearly.

In the model with covariates a statistically significant spatial autocorrelation is observed, with the Moran's I statistic equal to 0.1386 with a corresponding p-value for the null hypothesis of no spatial correlation of 0.04. In the model using no covariates, the Moran's I statistic equal to 0.1609 with a corresponding p-value for the null hypothesis of no spatial correlation of 0.02.

4.6.2 Likelihood and Autocorrelation models

The model used is the BYM model as presented in Section 4.2.

For the Portuguese alcohol abuse data the *response* variable is the number of cases of lifetime alcohol abuse per 100 inhabitants [2] and the covariates included are: proportion of population aged 18 to 34, proportion of population that is a regular alcohol user (as measured by Balsa et al. [4]), and proportion of males. The first two covariates are modelled with a natural cubic spline (3 degrees of freedom) and the third is modelled linearly.

For the Scotland lip cancer data the model covariate included is AFF modelled linearly.

[2] Package CARBayes version 4.3 can already handle missing values at the response variable level.

4.6.3 Matrices

Three different matrices were used, (a) the well-known and mostly used adjacency-based GMRF **W** matrix, (b) the distance-based GMRF **D** matrix (see Section 4.2 for both definitions) and (c) the application of the similarirty-based GRF **S** matrix defined in Section 4.3.

4.6.3.1 Portuguese alcohol abuse data

The relevant disease determinant factor considered for the **S**-based GRF model is the proportion of regular users of alcohol in each area unit, as collected by Balsa et al. [4]. Data are presented in Table 44, page 115. The rationale is that two areas with the "same"/close proportion of its population regularly consuming alcohol are more alike than two arbitrary areas. We use the proportion of alcohol users because Rose and Day [75] show that the number of alcohol abuse cases can be predicted by the number of alcohol use cases within a population. Therefore, the proportion of alcohol use cases within a population should be a good measure to use in defining similarity among areas.

The proportion of alcohol users is available by district and not by NUTS3. We are in the presence of misaligned spatial data [27], districts and NUTS3 are different partitions of the country, aggregating counties differently. We use a simple area interpolation approach and assume that alcohol users are distributed evenly throughout the district.

Model (a) **W**-based GMRF and model (b) **D**-based GMRF, are computed with the three covariates mentioned at the end of the previous subsection, while model (c) **S**-based GRF can include only two of the covariates, as the covariate proportion of regular alcohol users is already being used to build the similarity-based matrix.

4.6.3.2 Scotland lip cancer data

The matrix (c) **S**-based GRF is based on the available data, the AFF. We acknowledge that we do not have enough evidence that this factor can be considered as a determinant factor, but we will use it as another example of the new similarity-based GRF matrix. Model (a) **W**-based GMRF and model (b) **D**-based GMRF are computed with one covariate, the AFF, and the model (c) **S**-based GRF is computed with the intercept term only.

4.6.4 Inference

A fully Bayesian analysis of GMRF and/or GRF models is generally carried out using McMC, or more recently an approximate method using INLA, due to the intractable nature of posterior marginal distributions. In this case McMC was used. The analysis was implemented in R (version 3.2.2), with the package **CARBayes** [48]. The CARBayes

package uses a combination of Gibbs sampling and Metropolis steps. The variance parameters are Gibbs sampled from their full conditional truncated inverse gamma distributions, while the remaining parameters are updated using Metropolis steps with univariate random walk proposal distributions.

All analyses reported here implement a sum-to-zero constraint for the spatial random effects at each interaction of the McMC chain, and maintain a global intercept term in the linear predictor. This was done in order to solve the problem of identifiability of the model [10].

Posterior inference is based on 9 000 McMC samples, which are obtained by running one chain for 100 000 samples, by which convergence is assumed to have occurred. We ignore the first 10 000 samples as burn-in, and use the remaining 90 000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to avoid autocorrelation).

Pilot runs were carried out to establish appropriate burn-in using Geweke's diagnostic [29]. Convergence is assessed by visually monitoring trace and posterior density plots for each of the parameters.

4.6.5 Hyperpriors sensitivity tests

a. Portuguese Alcohol abuse model with $InverseGamma \sim (0.001, 0.001)$ [11]

   Using the BYM model with hyperpriors for both the structured and unstructured random effects as $IG \sim (0.001, 0.001)$ with the **S**-based GRF matrix model.

   For the first 10 attemps the model did not converge, meaning that the Geweke diagnostic [29] never was between $[-1.9, 1.9]$ for all parameters, including the fitted values (Y^*). On top of that the autocorrelation of the variances of the random effects were too high. See figures 4.2 and 4.3.

   Due to that, the hyperpriors needed to become more informative. After several attempts, the combination that worked the best was, for both parameters, the $IG \sim (0.1, 0.1)$. The autocorrelation problem was solved.

   The same process was followed for the **W**-based GMRF model (the adjacency-based GMRF matrix) and for the **D**-based GMRF model (the distance-based GMRF matrix). For the **W**-based GMRF model the hyperpriors needed to become more informative also ($IG \sim (0.1, 0.1)$), while for the **D**-based GMRF model the hyperpriors could remain at the non-informative level ($IG \sim (0.001, 0.001)$). For comparison reasons another model (d) was run with a $\mathbf{D}^*$-based GMRF matrix using the same weakly informative proper prior. Results for the (d) model are not reported because no changes were found in the posterior disease risks medians (between **D** and $\mathbf{D}^*$).

b. Portuguese Alcohol abuse model with $InverseGamma \sim (0.5, 0.005)$ [11]

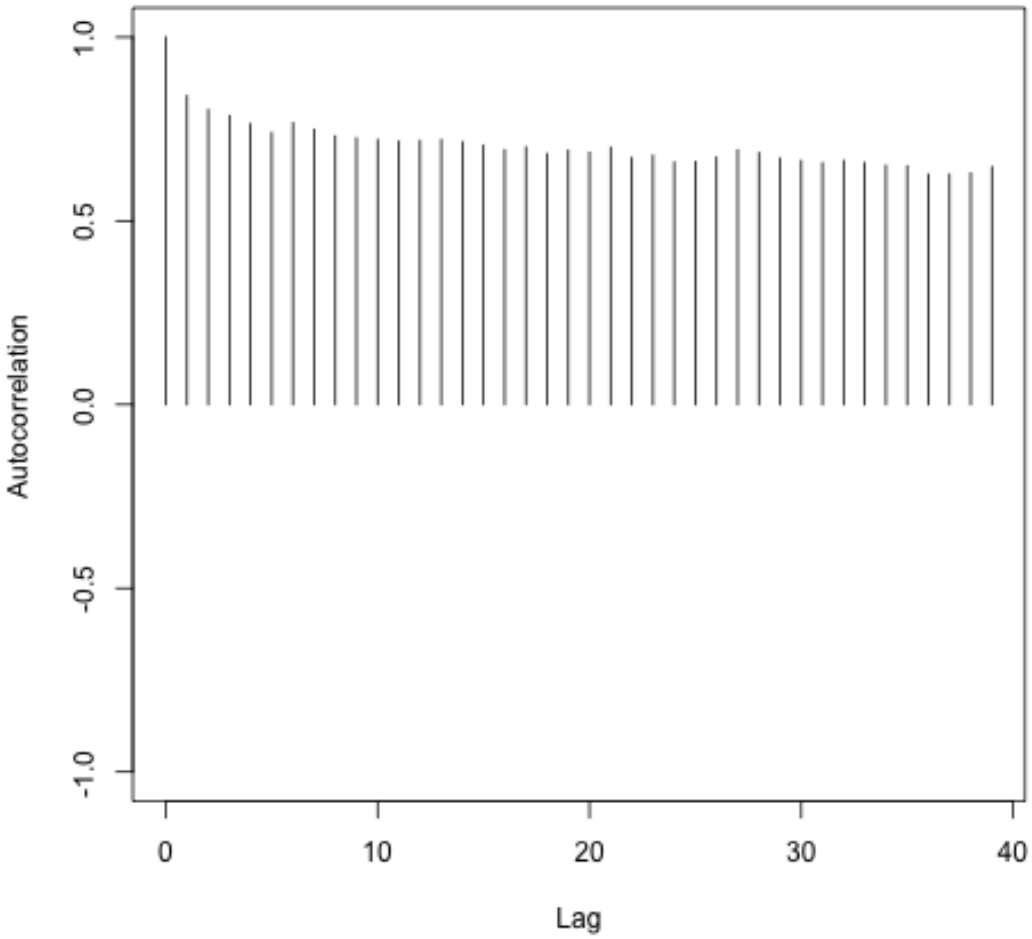

Figure 4.2: Variance for the structured random effects - autocorrelation.

Using the BYM model with hyperpriors for both the structured and unstructured random effects as $IG \sim (0.5, 0.005)$ with the **S**-based GRF matrix model.

For the first 50 attempts the model never converged and the autocorrelation of the variances of the random effects were too high. Making the hyperpriors more informative helped on the autocorrelation but not on the convergence. As we want to compare this model with the **W**-based GMRF and **D**-based GMRF models, we did not even try this hyperprior on those models.

c. Portuguese Alcohol abuse model with $U \sim (0, 10)$ [28]

Using the BYM model with hyperpriors for both structured and unstructured random effects as $U \sim (0, 10)$ with the **S**-based GRF matrix model.

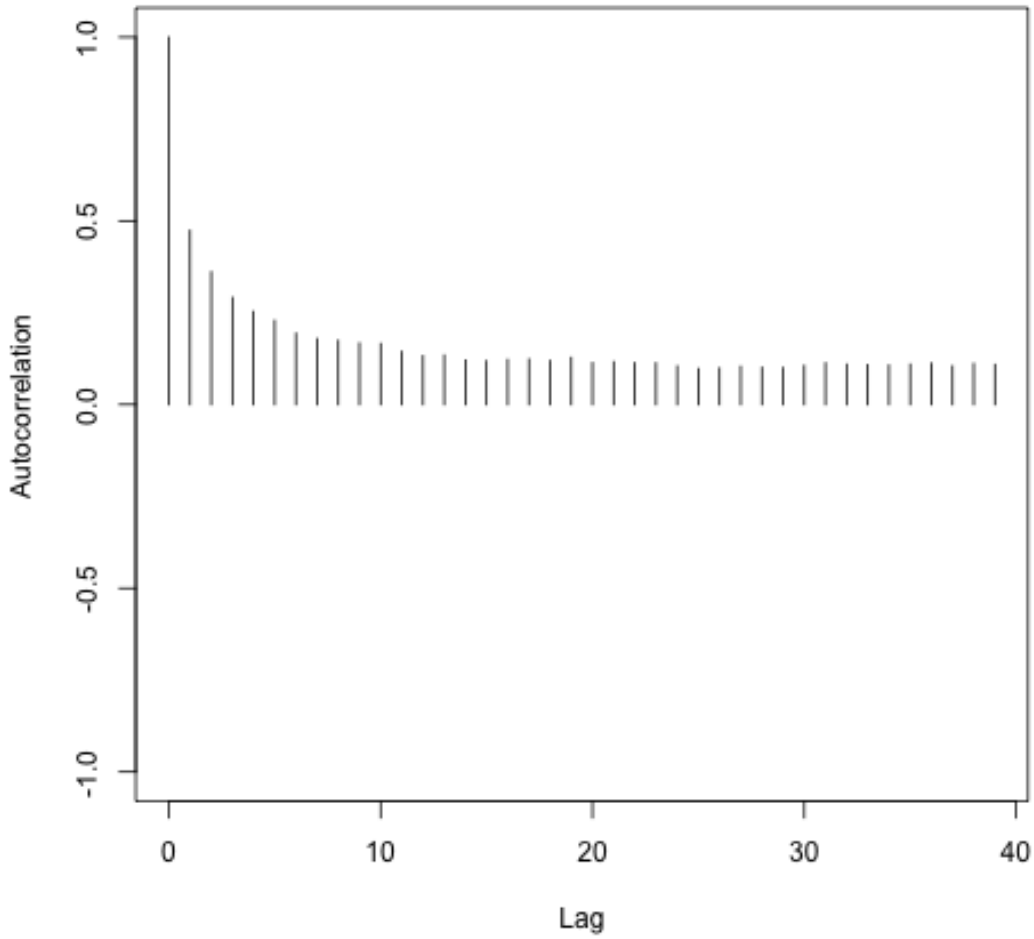

Figure 4.3: Variance for the unstructured random effects - autocorrelation.

The model converged and no autocorrelation problems appeared but the results, at the areas risk level, are somewhat different from those achieved both with the models run with the hyperpriors distributions as $InverseGamma$, and with the matrices type used.

As the Table 4.4 shows, the differences between the similarity-based GRF matrix and the adjacency-based GMRF matrix while using the $Uniform$ distribution are only at the low-risk area level. More areas are identified as low risk with the adjacency-based GRF matrix but when looking at results those are very close to the edge. In area number 3 the risk is between (0.3, 1.0), in area number 7 the risk is between (0.3, 1.0) and finally in area number 8 the risk is between (0.2, 0.9).

Hyperprior	Area risks	Area # Similarity-based matrix	Adjacency-based matrix
Uniform	High risk	21, 26	21, 26
Uniform	Low risk	2, 5, 10, 19	2, 3, 5, 7, 8, 10, 19
IG (0.1, 0.1)	High risk	16, 20, 21, 26	16, 20, 21, 26
IG (0.1, 01)	Low risk	2, 5	2, 5, 10, 19

Table 4.4: Results from the BYM model with the two types of matrices and two types of distributions. The risk is measured for 90% of the simulations.

The differences between the results achieved with the two different types of distributions happen at both levels, high and low risk. While comparing the high risk we see that the $InverseGamma$ results show two more areas. The results achieved with the $Uniform$ distribution for both areas are very close to 1, in the lower part of the interval, the area number 16 risk is between (0.9, 2.7) and the area number 20 risk is between (0.8, 2.5). At the low risk areas more differences are shown. The $Uniform$ distribution show many more areas but all of them with the superior limit of the interval at 1. For area number 3 the risk is between (0.3, 1.0), for area number 7 the risk is between (0.3, 1.0), for area number 8 the risk is between (0.2, 0.9), for area number 10 the risk is between (0.2, 0.9), and for area number 19 the risk is between (0.2, 1.0).

d. Conclusion

We decided to use the $IG \sim (0.1, 0.1)$ for the model with the **S**-based GRF matrix and for the model with the **W**-based GMRF matrix, and $IG \sim (0.001, 0.001)$ for the model with the **D**-based GMRF matrix. These hyperpriors distributions created more "prudent" results in our opinion (more high risk areas and fewer low risk areas). The low risk areas created by the $Uniform$ distribution models are too close to the edge, and that could lead to a situation where resources would be moved away from those areas while those were really needed.

As we used in the simulation study (subsection 4.4.1) exactly the same definitions used in the Alcohol abuse model, the hyperpriors chosen were the $IG \sim (0.001, 0.001)$ adjusted, as needed (more or less informative) to reach convergence and have acceptable levels of autocorrelation.

As the Scotland lip cancer data have been studied so many times and we wanted to be able to compare our results with those published we followed the hyperpriors already implemented and published. See, for example, [11], [79], [5] where only $InverseGamma$ distributions are used, and Ying MacNab has done extensive Bayesian sensitivity analysis on the BYM for the Scotland lip cancer data, which include commonly used hyperpriors such as $\tau \sim Gamma(0.01, 0.01)$, $\tau \sim Gamma(0.5, 0.005)$, $\sigma \sim$

$unif(0,10)$. Notable posterior sensitivity to hyperprior specifications with respect to the variance and spatial parameters were observed. However, modest posterior sensitivity was observed from the posterior prediction and inference for the relative risks. The results have not been published.

4.6.6 Prior and Hyperprior distributions

Prior distribution definition requires some care due to the use of weakly identifiable variables or high between-parameter posterior correlations. Variance components of the BYM model are not identifiable from the data, so identifiability of the individual effects (θ_i and ψ_i) is induced through the prior. Posterior inference needs to be based on informative hyperpriors, but it is often difficult to specify *a priori* the amount of structured similarity or unstructured heterogeneity expected in disease rates. In fact, this is one of the answers that models should provide, because this is of epidemiological interest, and hence strong prior distributions are to be avoided. Literature provides some suggestions on prior distributions to use. Best et al. [11] and the discussion therein proposes $InverseGamma$ distributions. Gelman [28] has some other considerations on this topic, specifically adding the $Uniform$ distribution to the possibilities. By using these two sources and after implementing several tests the more "prudent" solution (in terms of raised- and low-risk areas identification) seems to be the one implemented and explained above (see previous subsection).

A vague mean-zero Gaussian prior with variance 1 000 is specified for the regression parameters β (for the linear covariate) and for the intercept.

For the Portuguese alcohol abuse data, in order to obtain reasonable convergence properties and therefore reliable posterior estimates, some of the *a priori* distributions needed to become more informative. The *a priori* distributions for variances on both θ and ψ used are $InverseGamma \sim (0.001, 0.001)$ for the (b) **D**-based GMRF model, while models (c) **S**-based GRF and (a) **W**-based GMRF need a weakly informative proper prior on both parameters $InverseGamma \sim (0.1, 0.1)$. Because the class of $InverseGamma(\epsilon, \epsilon)$ priors are sensitive to the value of ϵ if the true variance is close to zero [28] another model (d) was run with a **D**-based GMRF matrix using the same weakly informative proper prior. Results for the (d) model are not reported because no changes were found in the posterior disease risks medians.

For the Scotland lip cancer data, the *a priori* distributions for variances on both θ and ψ are $InverseGamma \sim (0.001, 0.001)$. Only the model (c) **S**-based GRF needs a weakly informative proper prior on the structured spatial parameter ($InverseGamma \sim (0.1, 0.1)$). For the reasons mentioned above two more models (d) $\mathbf{W}^*$ (adjacency-based GMRF matrix model) and (e) $\mathbf{D}^*$ (distance-based GMRF matrix model) were run using the same weakly informative proper prior used to run the (c) **S**-based GRF matrix model. As the latter ones show minor differences in the posterior median disease risks when compared with (a)**W**-based GMRF and (b) **D**-based GMRF respectively, these are

Matrix	(a) W	(b) D	(c) S
DIC	136.1	135.8	135.0
p_D	19.1	18.2	17.8

Table 4.5: Portuguese alcohol abuse data - DIC results, which include the effective number of parameters in the model (p_D).

the ones used.

4.6.7 Edge effects

Based on the work of Lawson [45] we focused our analysis on the Portuguese areas close to the edges where there could be considerable distortion induced by missing *neighbours*. The average number of *neighbours* for the areas is 4.4 and we can find only two areas with significantly lower number of *neighbours*. One is in the North of the country (Minho which has only one *neighbour*), and could eventually have one more *neighbour*, the Spanish area of Galiza. The other one is in the South of the country (only have two *neighbours*), the area of Algarve, but that one could not have any more *neighbours*.

With the **D**-based GMRF matrix and the **S**-based GRF matrix model, specifically, that question is not so relevant because the matrix is built on distances, so all areas depend on all areas with different intensities. We have only one area that is disparate from all other areas in the case of the **S**-based GRF matrix model.

Therefore we have not implemented any method to deal with this question.

4.6.8 Models Results

4.6.8.1 Portuguese alcohol abuse data

Each model is assessed by the resulting Deviance Information Criterion (DIC) [79], in which a smaller value represents a better fitting model. Table 4.5 shows the results of the three models.

As DIC is a function of stochastic quantities generated under an McMC sampling scheme, it is subject to Monte Carlo sampling error. Whereas computing the precise standard errors for DIC values is a subject of on-going research [79], by running each model several times using different initial values of the parameters, randomly chosen, the DIC and p_D-estimates obtained never varied by more than 2. As such, and allowing for Monte Carlo error, all models seem (in terms of DIC performance) virtually indistinguishable in terms of the overall fit, and pragmatically, any of the models could be chosen.

Table 4.6 and Figure 4.4 (relationship between age, modelled with a natural cubic spline, and the number of alcohol abuse cases) show the posterior estimates under the

model (c) **S**-based GRF. It is worth mentioning that the goal of disease mapping is to estimate the pattern of disease risk over a geographical region and not to estimate associations between covariates and the disease cases. Nevertheless, due to the fact that coefficients were not found to be significantly different from zero (contrary to the expectations mentioned in Section 4.5), one must remember that this is an ecological study design, and the results must not be interpreted in terms of individual level cause and effect (the same results were found with the remaining two models). A possible explanation is ecological bias. The estimated residual random effects standard deviation for the BYM model (c) **S**-based GRF matrix are: a) the posterior sample median was 0.21 for the unstructured component (σ_θ^2) and b) 0.13 for the similarity structured component (σ_ψ^2), both the median posterior value and the wide intervals for both suggest a near split between the two components, which may result from the BYM identification issue [58].

Besides reporting and mapping the median posterior relative risk, the whole posterior distribution can be usefully exploited in an effort to detect true raised- and diminished-risk areas. Figure 4.5 shows the comparison between the posterior median disease risks obtained by the (c) **S**-based GRF model and the (a) **W**-based GMRF model in the top left corner. There is one difference that deserves attention. Model (a) **W**-based GMRF and model (b) **D**-based GMRF identify the area "Cova da Beira" as a diminished-risk area unlike the model (c) **S**-based GRF. Figure 4.6 shows on the left side the posterior probability of each area standardized morbidity ratio (SMR= Y_i/E_i)[7] being below 1, and on the right side the posterior probability of each area SMR being above 1, as produced by the (c) **S**-based GRF model, while the middle map shows the posterior median disease risks. Figure 4.7 and Figure 4.8 have exactly the same layout showing the results for the models with the (a) **W**-based GMRF and the (b) **D**-based GMRF matrices respectively.

Regarding "Cova da Beira", model (c) **S**-based GRF does not consider it as an area of diminished-risk because two out of three areas, which are more similar at the determinant factor level, have a crude SMR close to or above 1, with a SMR value equal to 0.57 with 90% of the simulations falling in the interval (0.28, 1.00). The (a) **W**-based GMRF and (b) **D**-based GMRF models are not able to overcome the fact that the crude SMR of the area (0.42) is very low, because its spatial *neighbours* have crude SMRs quite dispersed or missing (with the **W**-based GMRF SMR = 0.52 with 90% of the simulations falling in the interval (0.24, 0.96) and with the **D**-based GMRF SMR = 0.55 with 90% of the simulations falling in the interval (0.27, 0.99)). The result of the (c) **S**-based GRF matrix model seems most more prudent because the proportion of alcohol users in the area "Cova da Beira" is 64%, which is above the country mean value of 61%.

In the top right side Figure 4.5 compares the standard deviation values of the disease risks obtained by the (c) **S**-based GRF model and by the (a) **W**-based GMRF model. Overall, standard deviation values obtained by the (c) **S**-based GRF model are

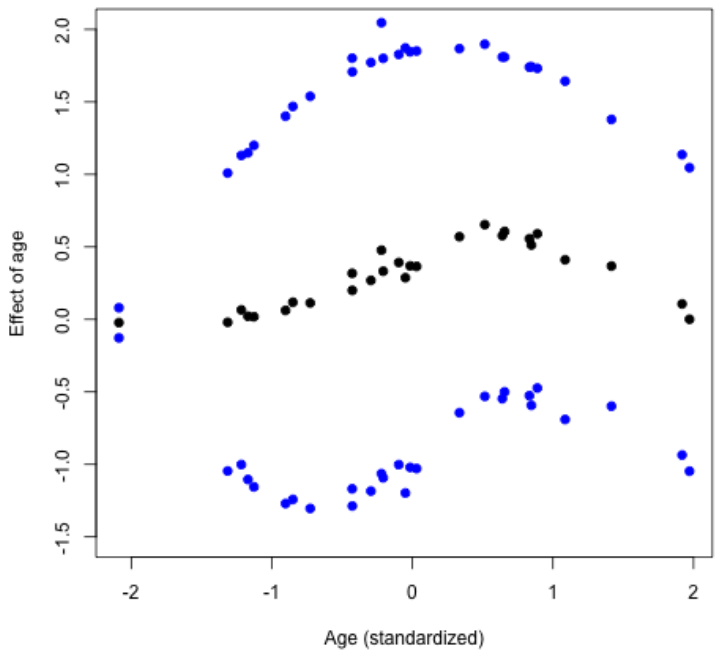

Figure 4.4: Portuguese alcohol abuse data - The estimated non-linear relationship between proportion of people aged 18 to 34 and the number of alcohol abuse cases. Blue curves delimit the 95% credible regions.

Parameter	**Prior distrib.**	**Prior mean**	**Prior std**	**McMC Post. median**	**2.5%**	**97.5%**
Intercept	$N(0, 10^3)$	0	10^3	1.61	-1.0	4.11
Male prop.	$N(0, 10^3)$	0	10^3	-0.07	-0.45	0.29
σ^2_ψ	IG (0.1, 0.1)	10^2	10^3	0.13	0.03	0.98
σ^2_θ	IG (0.1, 0.1)	10^2	10^3	0.21	0.05	0.70

Table 4.6: Portuguese alcohol abuse data - Model parameters summary for the model with (c) **S**-based GRF matrix.

smaller than those obtained by the (a) **W**-based GMRF model.

4.6.8.2 Scotland lip cancer data

The evaluation of the Scottish lip cancer model can be found in the literature [11]. It is worth mentioning that our results in terms of goodness-to-fit measures are, as expected, very close to those already published. Table 4.7 shows the results in terms

Matrix	(d) $\mathbf{W}^*$	(e) $\mathbf{D}^*$	(c) $\mathbf{S}$
DIC	298.2	309.5	308.4
p_D	32.8	39.1	39.3

Table 4.7: Scotland lip cancer data - DIC results, which include the effective number of parameters in the model (p_D).

of DIC and p_D-estimates for the three models. If we were choosing the model based on the DIC results, the model with the $\mathbf{W}^*$-based GMRF would be chosen. In terms of covariate coefficients, results for the models (a) $\mathbf{W}^*$-based GMRF and (b) $\mathbf{D}^*$-based GMRF are also consistent with those reported in the literature. We analyse the results only in terms of areas of raised- and diminished-risk.

Figure 4.5 shows the comparison between the posterior median disease risks obtained by the (c) $\mathbf{S}$-based GRF model and by the (d) $\mathbf{W}^*$-based GMRF model on the bottom left side, while the bottom right side of the figure shows the standard deviation values for the risk areas achieved by the same two models. In terms of the differences for the posterior median disease risks it can be said that the results achieved by the (a) $\mathbf{W}^*$-based GMRF model are more consistent with the published results while the differences on the standard deviation do not consistently favour either model.

4.7 Discussion

It has been shown by the case studies and by the simulation study that the similarity-based GRF model outperforms its spatial counterparts, the adjacency-based GMRF and the distance-based GMRF matrices, in correctly identifying raised- and diminished-risk areas in cases of no positive spatial correlation disease data. The case studies have also shown that enforcing an inappropriate spatial- or similarity-structure is likely to lead to poor risk estimates.

The decrease in the RMSE (and corresponding gain in efficiency) offered by this new matrix comes at the cost of a minor increase in bias; overall, however, the biases are still quite small. More complexity is added to the models, and the matrix is no longer defined by convenience or convention but needs to be based on data, which might not be available, be of poor quality or might be difficult to collect. Furthermore, it implies a deeper knowledge of the disease data at hand, and an extra effort to collect the determinant factors data.

McMC convergence with the similarity-based GRF models needs more informative hyperpriors, which may be a direct result of the quality of the data used, and more attention needs to be dedicated to this aspect. In the simulation study the determinant factor data create cases with few or no similar areas to borrow information from, and mechanisms to avoid this still need to be developed. This is important, because as in

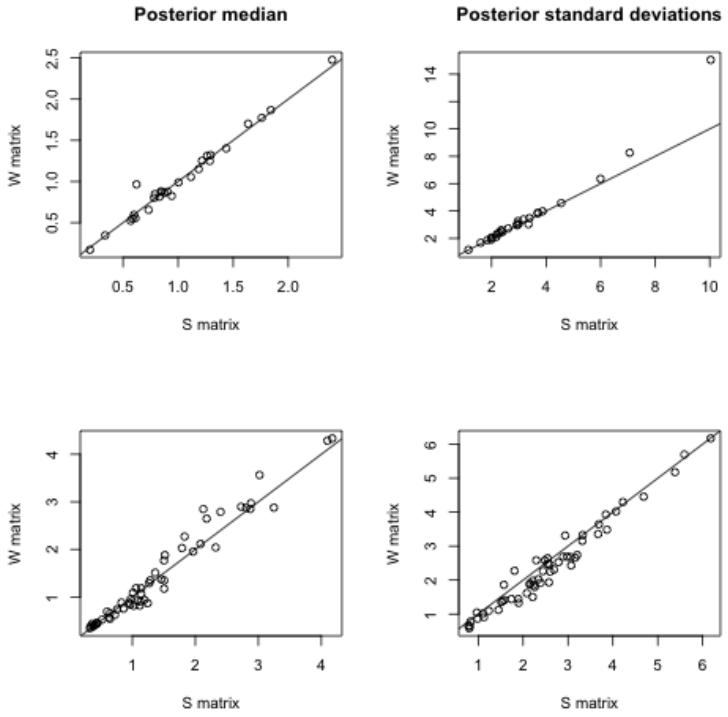

Figure 4.5: Top left: The posterior medians of the disease risks for the AAD in Portugal; Top right: The posterior standard deviation for the disease risks for the AAD in Portugal; Bottom left: The posterior medians of the disease risks for the Lip cancer in Scotland; Bottom right: The posterior standard deviation for the disease risks for the Lip cancer in Scotland.

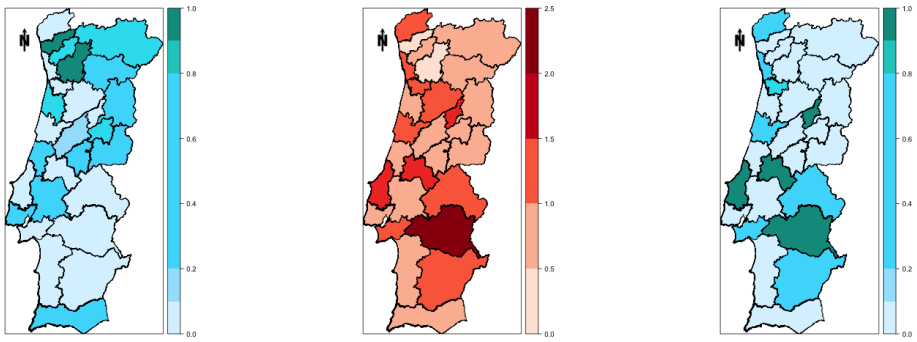

Figure 4.6: Portuguese AAD data - Results of the BYM model with the **S**-based GRF matrix - Left figure: map of posterior probabilities of SMR being below 1. Middle figure: map of the median posterior pattern of SMR. Right figure: map of posterior probabilities of SMR being above 1.

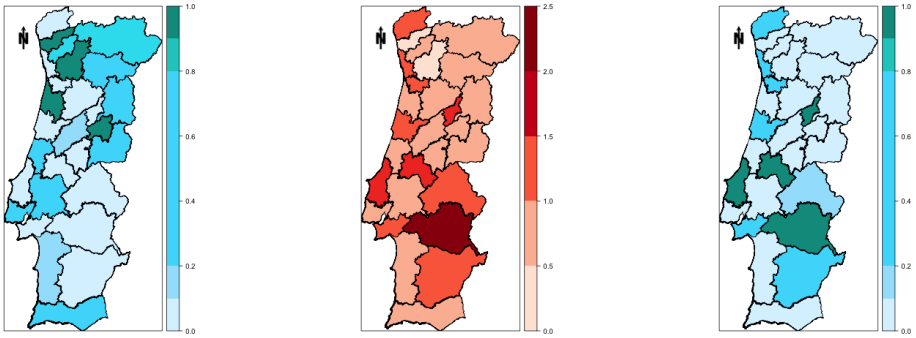

Figure 4.7: Portuguese AAD data - Results of the BYM model with the **W**-based GMRF matrix - Left figure: map of posterior probabilities of SMR being below 1. Middle figure: map of the median posterior pattern of SMR. Right figure: map of posterior probabilities of SMR being above 1.

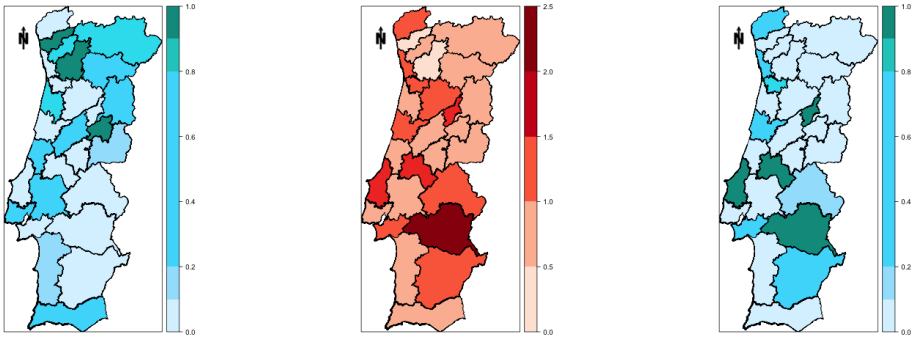

Figure 4.8: Portuguese AAD data - Results of the BYM model with the **D**-based GMRF matrix - Left figure: map of posterior probabilities of SMR being below 1. Middle figure: map of the median posterior pattern of SMR. Right figure: map of posterior probabilities of SMR being above 1.

the spatial setting, here too the "edge effects" can create bias in estimation and can lead to considerable increases in estimator variance at such locations and corresponding low reliability of estimation [45].

In Section 4.3 (equation 4.3), s_{ij} has been defined has a function of p_{ij}, the absolute gap between region i and region j in terms of disease determinant factor. However, p_{ij} can be defined in broader terms as the similarity between regions i and j. The similarity could correspond to the Euclidean distance in $\Re$ for p determinant factors:

$$p_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\prime(\mathbf{x}_i - \mathbf{x}_j)},$$

or even the multivariate version of the statistical distance, the Mahalanobis distance:

$$p_{ij} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\prime\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})},$$

where $\mathbf{x}_i\prime = (x_{1i}, x_{2i}, ..., x_{pi})$, $\mathbf{x}_j\prime = (x_{1j}, x_{2j}, ..., x_{pj})$, $\bar{\mathbf{x}}\prime = (\bar{x_1}, \bar{x_2}, ..., \bar{x_p})$ and $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix of the disease determinant factors. Further work must be done to check the impact of these distances in the GRF model.

Finally, our analysis of the Portuguese AAD reveals that four non-contiguous NUTS 3 localised in the Centre and South of the country show a raised risk. On the other side of the spectrum, only two areas can be considered as having a diminished risk, both in the North-west part of the country. The reasons why these regions show a high or a diminished risk remain unknown and further work is needed.

Chapter 5

Discussion

In this work we put together DM and epidemiology. DM has been a progressively expanding spatial statistics methodology for studies on spatial variations of disease and health outcomes in populations of small geographical regions and its associated ecological risk factors.

We begin by examining the existing Bayesian models for DM, such as those of Besag et al. [9] and Leroux et al. [55]. We examine their capabilities, drawbacks and developments. We also analyse the most recent developments in the field by showing the new LCAR model [52], which assumes local and not global smoothing, as do the BYM, MBYM and LLB. We present the MBYM model [59] and its advantages versus the BYM model. Furthermore, we detail the neighbourhood definition, as it is mostly known in the spatial statistics world today. We go a step further and include some actual definitions used on the geography science. Two conclusions can be drawn from this initial work: a) the weight matrix, used to define the neighbourhood structure while being recognized as pivotal has, surprisingly, received little attention to date; b) all of the parametric functions used in the weight matrix are related with the spatial locations of the small areas, either by adjacency or by distance. In summary, the use of the CAR model in the Bayesian DM context has been studied now for several years, and much is already known. Less attention has been given, so far, to the weight matrices and all have been inexorably linked to the spatial representation of the data.

This work covers the main epidemiology areas [66]. Starting by the analytical epidemiology, the design of the study (a cross-sectional study) is detailed in all its implementation aspects, highlighting the sampling method and the weighting. The weighting is especially important or the sample will not be representative of the population. Several types of bias are mentioned throughout Chapter 3, in order to clarify

what can and what cannot be concluded from the WMHSI data. Descriptive epidemiology is also covered with the calculation of the incidence and prevalence rates for AAD. This is an important part because it helps to understand the burden supported by the society and by the patients with AAD.

Multivariate data analysis methods are used to uncover the relationships between AAD and the demographic characteristics of the people in the sample, and between AAD and other mental disorders. The most up-to-date methods are applied, from multiplicative to additive models [91], to get the "real" picture of the disorder in Portugal. Understanding the AAD covariates is critical for the implementation of the DM models.

Calculation of prevalence and incidence rates in small areas and consequently SMRs requires age standardization methods and those are applied in accordance with the available data. Internal indirect standardization allows for the crude SMRs calculation in this case. DM models are GLMMs built on two basic components: a set of covariates and a set of random effects. The set of covariates used in subsection 3.5.2 are the ones that proved to be correlated with AAD at an individual level, age, and gender. The random effects element is at the core of the methodology, and is where the four models presented differ. The random effects are included to model any overdispersion and/or spatial correlation that might remain in the data after having being accounted for by the covariate information. Random effects are usually modeled by the class of CAR prior distributions, a type of GMRF. At this stage, the weight matrix is incorporated by means of an appropriate prior specification. In Chapter 3 the only weight matrix used is an adjacency matrix. For the BYM, LLB, and MBYM models a fixed first-order adjacency matrix, based on administratively defined areas, is used. For the LCAR model a random first-order adjacency matrix, based on administrative areas, and based on the spatial structure of the identified correlated mental disorders, is used.

The Deviance Information Criterion (DIC) can be considered as a Bayesian measure of goodness of fit or adequacy, penalized by an additional complexity term P_D [79]. The aim is to identify models that best explain the observed data, but with the expectation that they likely minimize uncertainty about observations generated in the same way. The MBYM model is the one achieving a lower DIC and therefore is the one chosen to explain the AAD risk distribution in Portugal. The mean value of $\lambda = 0.58$ achieved by the MBYM model indicates a local/global smoothing variance higher than unstructured variance. However, one of the drawbacks of the BYM, and consequently the MBYM models is their tendency to exhibit spurious geographical patterns in the area-specific risk estimates when no underlying excess risk is present [11]. A more in-depth analysis, using the measures of spatial association introduced in Chapter 2, reveals a statistically insignificant spatial autocorrelation present in the data.

In the presence of a set of data without positive spatial correlation, the "borrowing strength" mechanism that is actually present in the Bayesian hierarchical DM models

should not be used because it is based on the spatial weight matrix. Therefore, two options were available for the analysis of the AAD data: either proceed with the DM model omitting the structured random effect component, or change/replace the weight matrix structure. Consequently, the WMHSI offer us a valuable opportunity to explore options either than the actual weight matrix definition, because the data do not show a positive spatial correlation, meaning that the random effects representing the risk surface do not exhibit a global or even a local level of spatial smoothness.

As mentioned above, this work joins epidemiology and statistics, and from the epidemiology side it is evident that AAD is a disorder with determinant risk factors that do not vary systematically in space. Risk factors for AAD are intrinsic, deeply rooted in the Portuguese homogeneous culture. This fact creates a major challenge, because AAD does not have spatially varying associated ecological risk factors. The work done around the understanding of the disorder mechanisms provides us with the means to widen the scope of the existing CAR models. It is not possible to abuse alcohol without consuming it first, so using the proportion of alcohol users in each NUTS3 to measure how "likely" small areas would be, emerged as the best option to replace the spatial weight matrix. Furthermore, the data used to calculate the proportion of alcohol users in each NUTS3 is provided by a source not related with the WMHSI, avoiding inflating the possible measuring errors in the AAD cases data. The matrix becomes similarity-based, moving away from the adjacency- and distance-based ones.

This work proposes a GRF model with a similarity-based weight matrix in the conditional mean structure in cases of non-communicable diseases with intrinsic determinant factors because these are unlikely to vary systematically in space, differently from the extrinsic determinant factors of cancer and/or respiratory diseases. This similarity-based matrix enlarges the scope of existing CAR models, and to the best of our knowledge this is the first study to do that. A possible advantage is that in cases of disease data not exhibiting a positive spatial correlation the mechanism of borrowing strength of the GRF model can still be used to facilitate separation of systematic and random parts of the risk variation.

The similarity-based GRF matrix is therefore unique in its proposal of considering other sources of information, such as the disease determinant factors, and more investigation is needed before conclusions about its benefits and drawbacks can be advanced more concretely.

Therefore, our matrix is not based on a neighbourhood fashion, but is built in a way that gives more weight to those areas that are more alike, not because they are close to each other in space but because they are close with respect to a specific measure that determines the disease existence.

To test the adequacy of the proposal, a simulation study is conducted. The main goal of the simulation is to compare the results achieved with the BYM model with a) a GMRF adjacency-based matrix with the results achieved by the same model b) a GRF similarity-based matrix, for a disease with determinant risk factors not varying

systematically in space. In the a) case the matrix is built on spatial adjacencies and the determinant risk factor is included in the set of covariates. In the b) case the determinant risk factor information is used to build the matrix itself. On top of the simulation study the same approach is used in two data sets. Choosing one data set diametrically opposed in terms of spatial autocorrelation to the AAD helps to further understand the implications of the new model. The Scottish lip cancer data set has a statistically significant spatial autocorrelation, and the results obtained by the new GRF model lead to conclusions neither in line with those already published nor consistent with the information available. The Scottish example emphasizes the need for a good exploratory analysis of the data. The knowledge of the disorder etiology is crucial for the choice of the CAR model.

The results of both our simulation and our motivating example show that using a specific measure in the weight matrix is more efficient than using the same information as a covariate. Therefore, we consider that we are creating a new approach for the use of the matrix applied on the random effects.

The idea is not to replace the actual CAR use, but to enlarge it. We think that non-communicable diseases with determinant factors not positively spatially correlated would not benefit from the smoothing achieved by the actual CAR.

While starting and ending in DM, the work presented here, shows the importance of an integrated approach between epidemiology and statistics. Without the knowledge of the causes of AAD, this innovative way of looking at the problem would not have been possible.

There are many avenues for future work in this area. To start with, this model could also be implemented on the marginal structure, the marginal pair-wise correlations or covariances, not on the conditional structure, and this will mean, computationally, that different ways for the McMC implementation need to be found.

The hyperprior choice needs to be better studied to overcome the more than probable deficiencies of the data on the determinant risk factors. The discussion on informative and non-informative hyperpriors is very important due to the nature of posterior smoothing and rare events commonly seen in DM studies. However, in this particular case, the data at hand are not only disease cases but also the determinant risk factors which may not properly inform prior selection. While there are examples (such as the BYM model with the GMRF adjacency-based matrix in the Scotland lip cancer data) of modest posterior sensitivity to hyperprior specifications from the posterior prediction and inference for relative risks, the same cannot be simply assumed for any other model, especially when using the new GRF similarity-based matrix.

The "edge effects" is another particular area that needs to be expanded. A disparate area in terms of the determinant risk factor can create problem situations when that area has no information to "borrow strength" from. A kind of a "global random effect" potentially common to all areas might be a solution that would need to be implemented.

Furthermore, the methodology can be extended to the spatio-temporal domain and to the study of multiple diseases simultaneously. On the other hand there are plenty of undesired features in the BYM model [59], and therefore implementing the new GRF similarity-based matrix with another model, as the LLB model, for example, can provide other viable options.

Bibliography

[1] J. Alonso et al. "Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project." In: *Acta psychiatrica Scandinavica. Supplementum* 420 (2004), pp. 21–7.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition)*. 4th. Washigton, DC: American Psychiatic Association, 1994.

[3] L. Anselin. *Spatial Econometrics: Methods and Models*. New York: Kluwer Academic Publisher, 1988.

[4] C. Balsa, C. Vital, and L. Pascueiro. *O consumo de bebidas alcoólicas em Portugal Prevalências e padrões de consumo, 2001-2007*. Tech. rep. Lisbon: IDT; CesNova-Centro de estudos em sociologia, faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, 2011.

[5] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data (Second edition)*. Boca Raton: Chapman&Hall/CRC, 2014.

[6] H. Baptista, J. M. Mendes, Y. C. MacNab, M. Xavier, and J. M. C. de Almeida. "A Guassian random field model for similarity-based smoothing in Bayesian disease mapping". In: *To appear in Statistical methods in medical research* ().

[7] H. Baptista, J. M. Mendes, J. Caldas de Almeida, and M. Xavier. "Alcohol abuse disorder prevalence and its distribution across Portugal. A Disease mapping approach". In: *REVSTAT* (2015).

[8] J. Besag. "Spatial Interaction and the Statistical Analysis of Lattice Systems". In: *Journal of the Royal Statistical Society* 36.2 (1974), pp. 192–236.

[9] J. Besag, J. York, and A. Mollié. "Bayesian Image Restoration, with Two Applications in Spatial Statistics". In: *Annals of the Institute of Statistical Mathematics* 43.1 (1991), pp. 1–20.

[10] J. Besag, P. Green, D. Higdon, and K. Mengersen. "Bayesian computation and stochastic systems". In: *Statistical Science* 10.1 (1995), pp. 8–66.

[11] N. Best, R. Arnold, A. Thomas, L. Waller, and E. Conlon. "Bayesiann Models for Spatially Correlated Disease and Exposure Data". In: *Bayesian Statistics 6*. Ed. by J. Bernardo, J. Berger, A. Dawid, and A. Smith. Oxford: Oxford Science Publications, 1999, pp. 131–147.

[12] N. Best, S. Richardson, and A. Thomson. "A comparison of Bayesian spatial models for disease mapping." In: *Statistical methods in medical research* 14.1 (2005), pp. 35–59.

[13] E. Bromet et al. "Cross-national epidemiology of DSM-IV major depressive episode." In: *BMC medicine* 9.1 (2011), p. 90.

[14] C. H. Brown. "Analyzing preventive trials with generalized additive models". In: *American Journal of Community Psychology* 21.5 (1993), pp. 635–664.

[15] D. Clayton and J. Kaldor. "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping." In: *Biometrics* 43.3 (1987), pp. 671–81.

[16] P. S. Coelho and L. N. Pereira. "A Spatial Unit Level Model For Small Area Estimation". In: *REVSTAT* 9.2 (2011), pp. 155–180.

[17] P. Congdon. "A spatially adaptive conditional autoregressive prior for area health data". In: *Statistical Methodology* 5.6 (2008), pp. 552–563.

[18] J. P. Connor, P. S. Haber, and W. D. Hall. "Alcohol use disorders". In: *The Lancet* 6736 (2015), p. 122.

[19] N. A. Cressie and N. H. Chan. "Spatial Modeling of Regional Variables". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 393–401.

[20] C. B. Dean, M. D. Ugarte, and a. F. Militino. "Detecting interaction between random region and fixed age effects in disease mapping." In: *Biometrics* 57.1 (2001), pp. 197–202.

[21] L. Degenhardt et al. "Toward a global view of alcohol, tobacco, cannabis, and cocaine use: findings from the WHO World Mental Health Surveys." In: *PLoS medicine* 5.7 (2008), e141.

[22] K. Demyttenaere, R. Bruffaerts, J. Posada-Villa, I. Gasquet, V. Kovess, J. P. Lepine, M. C. Angermeyer, S. Bernert, G. de Girolamo, P. Morosini, G. Polidor, and S. C. "Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys". In: *JAMA : the journal of the American Medical Association* 291.21 (2004), pp. 2581–2590.

[23] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Second. London: Arnold, a member of the Hodder Headline Group, 2003.

[24] A. Earnest, G. Morgan, K. Mengersen, L. Ryan, R. Summerhayes, and J. Beard. "Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models." In: *International journal of health geographics* 6 (2007), p. 54.

[25] P. B. English, M. Kharrazi, S. Davies, R. Scalf, L. Waller, and R. Neutra. "Changes in the spatial pattern of low birth weight in a southern California county: The role of individual and neighborhood level factors". In: *Social Science and Medicine* 56.10 (2003), pp. 2073–2088.

[26] A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression - the analysis of spatially relationships*. West Sussex: John Wiley & Sons, Ltd., 2002.

[27] A. E. Gelfand. "Misaligned Spatial Data: The Change of Support Problem". In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp. Boca Raton: Taylor & Francis Group, 2010. Chap. 29, pp. 517–539.

[28] A. Gelman. "Prior distribution for variance parameters in hierarchical models". In: *Bayesian Analysis* 1.3 (2006), pp. 515–533.

[29] J. Geweke. "Evaluating the accurary of sampling-based approaches to the calculation of posterior moments". In: *Bayesian Statistics*. Oxford: Oxford University Press, 1992, pp. 169–193.

[30] M. D. Glantz et al. "Alcohol abuse in developed and developing countries in the World Mental Health Surveys: Socially defined consequences or psychiatric disorder?" In: *The American journal on addictions / American Academy of Psychiatrists in Alcoholism and Addictions* 23.2 (2014), pp. 145–55.

[31] P. Goovaerts and S. Gebreab. "How does Poisson kriging compare to the popular BYM model for mapping disease risks?" In: *International journal of health geographics* 7 (2008), p. 6.

[32] D. A. Griffith. "Some guidelines for specifying the geographic weights matrix contained in spatial statistical models." In: *Practical Handbook of Spatial Statistics*. Ed. by S. L. Arlinghaus. Boca Raton: CRC Press, 1996, pp. 65–82.

[33] J. M. Haro, S. Arbabzadeh-bouchez, T. S. Brugha, G. De, M. E. Guyer, R. Jin, J. P. Lepine, F. Mazzi, B. Reneses, G. Vilagut, N. A. Sampson, and R. C. Kessler. "Concordance of the Composite International with standardized clinical assessments in the WHO World Mental Health Surveys". In: 15.4 (2006), pp. 167–180.

[34] T. Hastie and R. Tibshirani. "Generalized Additive Models". In: *Statistical Science* 1.3 (1986), pp. 297–318.

[35] L. Held and H. Rue. "Conditional and intrinsic autoregressions". In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp. Boca Raton: Taylor & Francis Group, 2010. Chap. 13, pp. 201–215.

[36] J. S. Hodges and B. J. Reich. "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love". In: *The American Statistician* 64.4 (2010), pp. 325–334.

[37] INE. *Censos 2001: resultados definitivos*. Tech. rep. INE - Statistics Portugal, 2001.

[38] INE. *Portugal: Census 2011*. Tech. rep. INE - Statistics Portugal, 2011.

[39] A. Kalaydjian, J. Swendsen, W.-T. Chiu, L. Dierker, L. Degenhardt, M. Glantz, K. R. Merikangas, N. Sampson, and R. Kessler. "Sociodemographic predictors of transitions across stages of alcohol use, disorders, and remission in the National Comorbidity Survey Replication." In: *Comprehensive psychiatry* 50.4 (2009), pp. 299–306.

[40] R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H.-U. Wittchen, and K. S. Kendler. "Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States Results From the National Comorbidity Survey". In: *JAMA Psychiatry* 51.1 (1994), pp. 8–19.

[41] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters. "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication." In: *Archives of general psychiatry* 62.6 (2005), pp. 593–602.

[42] M. King, I. Nazareth, G. Levy, C. Walker, R. Morris, S. Weich, J. A. Bellón-Saameño, B. Moreno, I. Svab, D. Rotar, J Rifel, H.-I. Maaroos, A. Aluoja, R. Kalda, J. Neeleman, M. I. Geerlings, M. Xavier, M. C. de Almeida, B. Correa, and F. Torres-Gonzalez. "Prevalence of common mental disorders in general practice attendees across Europe." In: *The British journal of psychiatry : the journal of mental science* 192.5 (2008), pp. 362–7.

[43] A. Kleinman. "Global mental health: a failure of humanity". In: *The Lancet* 374.9690 (2009), pp. 603–604.

[44] V. Kovess-Masféty, D. Wiersma, M. Xavier, J. M. C. de Almeida, M. G. Carta, J. Dubuis, E. Lacalmontie, J. Pellet, J.-L. Roelandt, F. Torres-Gonzalez, B. Moreno Kustner, and D. Walsh. "Needs for care among patients with schizophrenia in six European countries: a one-year follow-up study." In: *Clinical practice and epidemiology in mental health : CP & EMH* 2 (2006), p. 22.

[45] A. B. Lawson. *Bayesian Disease Mapping Hierarchical modeling in Spatial epidemiology*. Ed. by N. Keiding, B. Morgan, C. Wikle, and P. van der Heijden. Boca Raton: Chapman and Hall Book, 2009.

[46] A. B. Lawson, A. Biggeri, D. Boehning, E. Lesaffre, J.-F. Viel, A. Clark, P. Schlattmann, and F. Divino. "Disease mapping models : an empirical evaluation". In: *Statistics in Medicine* 19.19 (2000), pp. 2217–2241.

[47] D. Lee. "A comparison of conditional autoregressive models used in Bayesian disease mapping". In: *Spatial and spatio-temporal epidemiology* 2 (2011), pp. 79–89.

[48] D. Lee. "CARBayes : An R Package for Bayesian Spatial". In: *Journal of Statistical Software* 55.13 (2013), pp. 1–24.

[49] D. Lee and R. Mitchell. "Boundary detection in disease mapping studies". In: *Biostatistics (Oxford, England)* 13.3 (2012), pp. 415–426.

[50] D. Lee and R. Mitchell. "Locally adaptive spatial smoothing using conditional auto-regressive models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62.4 (2013), pp. 593–608.

[51] D. Lee and R. Mitchell. "Controlling for localised spatio-temporal autocorrelation in long-term air pollution and health studies." In: *Statistical methods in medical research* 23.6 (2014), pp. 488–506.

[52] D. Lee, A. Rushworth, and S. K. Sahu. "A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution." In: *Biometrics* 70.2 (2014), pp. 419–29.

[53] S. Lee, W.-J. Guo, A. Tsang, Y.-L. He, Y.-Q. Huang, M.-Y. Zhang, Z.-R. Liu, Y.-C. Shen, and R. C. Kessler. "Associations of cohort and socio-demographic correlates with transitions from alcohol use to disorders and remission in metropolitan China." In: *Addiction (Abingdon, England)* 104.8 (2009), pp. 1313–23.

[54] A. S. Leoussi, ed. *Encyclopaedia of Nationalism*. New Brunswick, New Jersey: Transaction Publishers, 2001, p. 1.

[55] B. G. Leroux, X. Lei, and N. Breslow. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence". In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Ed. by M. E. Halloran and D. Berry. Vol. 116. The IMA Volumes in Mathematics and its Applications. New York, NY: Springer New York, 2000, pp. 179–191.

[56] B. Lu, M. Charlton, and A. S. Fotheringham. "Geographically Weighted Regression using a non-Euclidean distance metric with a study on London house price data". In: *International Journal of Geographical Information Science* 28.4 (2014), pp. 660–681.

[57] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. "The BUGS project: Evolution, critique and future directions". In: *Statistics in medicine* 28 (2009), pp. 3049–3067.

[58] Y. C. MacNab, S. Read, M. Strong, T. Pearson, R. Maheswaran, and E. Goyder. "Bayesian hierarchical modelling of noisy spatial rates on a modestly large and discontinuous irregular lattice". In: *Statistical Methods in Medical Research* 23.6 (2014), pp. 552–571.

[59] Y. C. MacNab. "On Gaussian Markov random fields and Bayesian disease mapping." In: *Statistical methods in medical research* 20.1 (2011), pp. 49–68.

[60] Y. C. MacNab. "On identification in Bayesian disease mapping and ecological-spatial regression models." In: *Statistical methods in medical research* 23.2 (2014), pp. 134–55.

[61] P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1989, p. 532.

[62] P. Michael King, MD, P. Carl Walker, BSc, M. Gus Levy, P. Christian Bottomley, D. Patrick Royston, D. Scott Weich, MBBS, P. Juan Ángel Bellón-Saameño, MD, P. Berta Moreno, P. Igor Švab, MD, M. Danica Rotar, MD, M. J. Rifel, and P. Heidi-Ingri. "Development and Validation of an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice Attendees The PredictD Study". In: *JAMA Psychiatry* 65.12 (2008), pp. 1368–1376.

[63] P. A. P. Moran. "Notes on Continuous Stochastic Phenomena". In: *Biometrika* 1.2 (1950), pp. 17–23.

[64] National Mental Health Development Unit. *The costs of mental ill health*. Tech. rep. London, UK: NMHDU, 2010. URL: www.nmhdu.org.uk.

[65] Y. D. Neumark, C. Lopez-Quintero, A. Grinshpoon, and D. Levinson. "Alcohol drinking patterns and prevalence of alcohol-abuse and dependence in the Israel National Health Survey." In: *The Israel journal of psychiatry and related sciences* 44.2 (2007), pp. 126–35.

[66] J. Olsen, K. Christensen, J. Murray, and A. Ekbom. *An Introduction To Epidemiology for Health professionals*. Ed. by W. Ahrens. Springer Science+Business Media, 2010. ISBN: 9781441914965. DOI: 10.1007/978-1-4419-1497-2.

[67] V. Patel. "Global mental health: from science to action." In: *Harvard review of psychiatry* 20.1 (2012), pp. 6–12.

[68] B.-E. Pennell, A. Mneimneh, Zeina N. Bowers, S. Chardoul, J. E. Wells, M. C. Viana, D. Karl, N. Gebler, S. Florescu, Y. He, Y. Huang, T. Tomov, and G. V. Saiz. "No Title". In: *The WHO World Mental Health Surveys*. Ed. by R. Kessler and T. Ustun. New York, USA: Cambridge University Press, 2008, pp. 33–57.

[69] A. E. Raftery and J. D. Banfield. "Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics (Discussion of Besag et al.)" In: *Annals of the Institute of Statistical Mathematics* 43.1 (1991), pp. 32–43.

[70] J. Rao. "Small area estimation: Methods and applications". In: *Applications of small area estimation techniques in the social sciences*. Mexico City, 2012.

[71] D. A. Regier, M. E. Farmer, D. S. Rae, J. K. Myers, M. Kramer, L. N. Robins, L. K. George, M. Karno, and B. Z. Locke. "One-month prevalence of mental disorders in the United States and sociodemographic characteristics: the Epidemiologic Catchment Area study". In: *Acta Psychiatrica Scandinavica* 88.1 (1993), pp. 35–47.

[72] W. H. Report. "The World Health Report 2004 - changing history". In: *World Health* 95.1 (2004), 96p. URL: http://www.who.int/whr/2004/en/index.html.

[73] S. Richardson, A. Thomson, N. Best, and P. Elliott. "Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies". In: *Environmental Health Perspectives* 112.9 (2004), pp. 1016–1025.

[74] B. Ripley. *Spatial Statistics*. New Jersey: John Wiley & Sons, Inc.New Jersey, 1981.

[75] G. Rose and S. Day. "The population mean predicts the number of deviant individuals." In: *BMJ (Clinical research ed.)* 301.6759 (1990), pp. 1031–1034.

[76] H. Rue, S. Martino, and N. Chopin. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2 (2009), pp. 319–392.

[77] D. Sanco. *The Mental Health Status of the European Population*. Tech. rep. April. EORG, 2003.

[78] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. a. Ghali. "Comparison of distance measures in spatial analytical modeling for health service planning." In: *BMC health services research* 9 (2009), p. 200. ISSN: 1472-6963.

[79] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. D. Linde. "Bayesian measures of model complexity and fit". In: *Journal of the Royal Society Statistical Society. Series B (Statistical Methodology)* 64.4 (2002) (2010), pp. 583–639.

[80] Statistics Netherlands. *Blaise Developer's Guide*. 1999.

[81] J. Swendsen, K. P. Conway, L. Degenhardt, L. Dierker, M. Glantz, R. Jin, K. R. Merikangas, N. Sampson, and R. C. Kessler. "Socio-demographic risk factors for alcohol and drug dependence: the 10-year follow-up of the national comorbidity survey". In: *Addiction* 104.8 (2009), pp. 1346–1355.

[82] J. Wakefield. "Disease mapping and spatial regression with count data." In: *Biostatistics (Oxford, England)* 8.2 (2007), pp. 158–83.

[83] J. Wakefield and H. Lyons. "Spatial Aggregation and the ecological fallacy". In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp. Boca Raton: Taylor & Francis Group, 2010. Chap. 30, pp. 541–558.

[84] L. A. Waller and B. P. Carlin. "Disease Mapping". In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp. Boca Raton: Taylor & Francis Group, 2010. Chap. 14, pp. 217–243.

[85] L. A. Waller and C. A. Gotway. *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Inc., 2004.

[86] S. Weich and R. Araya. "International and regional variation in the prevalence of common mental disorders: Do we need more surveys?" In: *British Journal of Psychiatry* 184.APR. (2004), pp. 289–290.

[87] WHO. "Neuroscience of Psychoactive Substance Use and Dependence". In: *World Health Organization, Geneva* 99.10 (2004), pp. 1361–1362.

[88] WHO. *Global Status Report on alcohol and health*. Tech. rep. WHO, 2014.

[89] H. U. Wittchen and F. Jacobi. "Size and burden of mental disorders in Europe - A critical review and appraisal of 27 studies". In: *European Neuropsychopharmacology* 15.4 (2005), pp. 357–376.

[90] H. Wittchen, F. Jacobi, J. Rehm, A. Gustavsson, M. Svensson, B. Jönsson, J. Olesen, C. Allgulander, J. Alonso, C. Faravelli, L. Fratiglioni, P. Jennum, R. Lieb, A. Maercker, J. van Os, M. Preisig, L. Salvador-Carulla, R. Simon, and H.-C. Steinhausen. "The size and burden of mental disorders and other disorders of the brain in Europe 2010". In: *European Neuropsychopharmacology* 21.9 (2011), pp. 655–679.

[91] S. N. Wood. *Generalized Additive Models: an introduction with R*. Chapman and Hall/CRC Press, 2006.

[92] World Health Organisation. *Mental Health: Facing the Challenges, building solutions*. 2005, p. 195.

[93] World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders*. 1993.

[94] M. Xavier, H. Baptista, J. M. Mendes, P. Magalhães, and J. M. Caldas-de Almeida. "Implementing the World Mental Health Survey Initiative in Portugal - rationale, design and fieldwork procedures." In: *International journal of mental health systems* 7.1 (2013), pp. 7–19.

Appendix A

R codes

This appendix shows the R codes used in Chapter 4.

A.1 Simulation study

```
require(MASS)
library(shapefiles)
library(sp)
library(spdep)
library(spam)
library(truncdist)
library(coda)
library(CARBayes)

################################################
#     Spatial Neighbourhood Matrix
################################################
SA<-read.table('SA.txt',header=TRUE)
SA<-as.matrix (SA)

################################################
#      Spatial Distances Matrix
################################################
y <- 1:100
x <- 1:100
grd<-data.frame(x,y)
```

```
D<-matrix(0,100,100)
for (i in 1:100){
for (j in 1:100){
D[i,j]<-sqrt((grd$x[i]-grd$x[j])^2+(grd$y[i]-grd$y[j])^2)
}}

zeta<- -mean(D)/log(0.01) # Best et al (1999), p. 136
D<-exp(-D/zeta)
for (i in 1:100){D[i,i]<-mean(D[i,-i])}


################################################
#        Simulation => Observed
################################################

### Region (lattice)
y <- 1:10
x <- 1:10
grd <- expand.grid(y,x);names(grd)<-c("y","x")
grd$u<-c(runif(100,0,1))


#### Expected cases
grd$E <- rep(40,100)

#### Covariates
grd$x1<- rnorm(100)
grd$x2<- rnorm(100)


Sigma<-matrix(0,100,100)
diag(Sigma)<-0.8*rgamma(100,shape=1,scale=1)
#Independ. 0, Moderate 0.5, Strong 0.8
for (i in 1:99){
for (j in (i+1):100){
Sigma[i,j]<-0.8*sqrt(Sigma[i,i])*sqrt(Sigma[j,j])
Sigma[j,i]<-0.8*sqrt(Sigma[i,i])*sqrt(Sigma[j,j])
}}

### Disease determinant (dd)
```

```
grd$dd<-mvrnorm(1, mu=rep(0,100), Sigma)
grd$dd<-grd$dd[order(grd$u)]

### "Structured" spatial process
grd$phi<-0.9*grd$dd+rnorm(100,0,0.05)

# Correlation between disease determinant and spatial process
cor(grd$dd,grd$phi)

#### Unstructured spatial process
grd$theta <- rnorm(100, sd=0.2)

#### Relative risk
grd$RR <- exp(-0.2 +  0.1*grd$x1 + 0.1*grd$x2 + grd$theta + grd$phi)

#### Generating cases
grd$Y <- rpois(100, lambda=grd$E*grd$RR)
summary(grd$Y)

# Correlation between cases and disease determinant
cor(grd$Y,grd$dd)

################################################
#      Similarity Distances Matrix
################################################
S <-matrix(0,100,100)
for(i in 1:100){
for(j in 1:100){
S[as.numeric(row.names(grd)[i]),as.numeric(row.names(grd)[j])]
<- abs(grd$dd[i]-grd$dd[j])
}
}
S<-exp(-S/(-mean(S)/log(0.01)))
for (i in 1:100){S[i,i]<-mean(S[i,-i])}

################################################
#      Run GLMM - BYM models
################################################
#Run model on similarity
formS<-Y~x1+x2+offset(E)
modelSS<- S.CARbym(formula=formS, family="poisson", data=grd,
```

```
W=S,burnin=10000,n.sample=100000,thin=10,prior.tau2=c(0.001,0.001),
prior.sigma2=c(0.01,0.01))
modelSS

Sfittedt<-summarise.samples(modelSS$samples$fitted,
quantiles=c(0.5, 0.025, 0.975))
Sfitted<-Sfittedt$quantiles[,1]
Sfittedl<-Sfittedt$quantiles[,2]
Sfittedh<-Sfittedt$quantiles[,3]

#Run model on adjacency
formA<-Y~x1+x2+dd+offset(E)
modelSA<- S.CARbym(formula=formA, family="poisson", data=grd,
W=SA,burnin=10000,n.sample=100000,thin=10,prior.tau2=c(0.1,.1),
prior.sigma2=c(0.01,0.01))
modelSA

Afittedt<-summarise.samples(modelSA$samples$fitted,
quantiles=c(0.5, 0.025, 0.975))
Afitted<-Afittedt$quantiles[,1]
Afittedl<-Afittedt$quantiles[,2]
Afittedh<-Afittedt$quantiles[,3]

#Run model on distance
modelSD<- S.CARbym(formula=formA, family="poisson", data=grd,
W=D,burnin=10000,n.sample=100000,thin=10,prior.tau2=c(0.1,.1),
prior.sigma2=c(0.01,0.01))
modelSD

Dfittedt<-summarise.samples(modelSD$samples$fitted,
quantiles=c(0.5, 0.025, 0.975))
Dfitted<-Dfittedt$quantiles[,1]
Dfittedl<-Dfittedt$quantiles[,2]
Dfittedh<-Dfittedt$quantiles[,3]


#########################################################
#      Calculate RMSE, Bias and Coverage & Retain result
#########################################################
strong<-matrix(0,200,6)
strongcov_A<-matrix(0,200,100)
```

```
strongcov_S<-matrix(0,200,100)

#RMSE and Bias
RMSE_A<-rmse(Afitted, grd$Y)
RMSE_S<-rmse(Sfitted, grd$Y)
RMSE_D<-rmse(Dfitted, grd$Y)
BIAS_A<-pbias(Afitted, grd$Y)
BIAS_S<-pbias(Sfitted, grd$Y)
BIAS_D<-pbias(Dfitted, grd$Y)
Coverage_A<-ifelse(grd$Y>=Afittedl & grd$Y<= Afittedh, 1, 0)
Coverage_S<-ifelse(grd$Y>=Sfittedl & grd$Y<= Sfittedh, 1, 0)
Coverage_D<-ifelse(grd$Y>=Dfittedl & grd$Y<= Dfittedh, 1, 0)

strong[500,]=c(RMSE_A,RMSE_S,BIAS_A,BIAS_S, RMSE_D,BIAS_D)
strongcov_A[500,]=c(Coverage_A)
strongcov_S[500,]=c(Coverage_S)
strongcov_D[100,]=c(Coverage_D)

strongcov_A[501,]<-colSums(strongcov_A)/5
strongcov_S[501,]<-colSums(strongcov_S)/5
strongcov_D[101,]<-colSums(strongcov_D)

mean(strong[1:500,1])
mean(strong[1:500,2])
mean(strong[1:500,3])
mean(strong[1:500,4])
summary(t(strongcov_A[501,]))
summary(t(strongcov_S[501,]))
mean(strongD[1:100,1])
mean(strongD[1:100,2])
summary(t(strongcov_D[101,]))
```

A.2 DM models

```
library(shapefiles)
library(sp)
library(spdep)
library(spam)
library(truncdist)
library(coda)
library(CARBayes)
```

```
library(splines)

################################################
#    Read all needed data and create Matrices
################################################
data<-read.table('data.txt',header=TRUE)
data$cases_alah_na<-ifelse(data$cases_alah==0,NA,data$cases_alah)
data$cases_alah_na_00<-round(data$cases_alah_na/data$population*100,0)
data$SMR_alah<-data$cases_alah_na/data$alah_expected
data$expected_00<-round(data$alah_expected/data$population*100,1)
data$Balsapc<-scale(data$Balsa,center=TRUE,scale=TRUE)

#Create Distance matrix, Best et. al. (1999)#
n<-length(data$XX)
matrix<-matrix(NA,nrow=28,ncol=28)
for (i in 1:n){
for (k in i:n){
matrix[i,k]<-sqrt((data$XX[i]-data$XX[k])^2+(data$YY[i]-data$YY[k])^2)
}
}
Dis<-forceSymmetric(matrix)
Dis<-as.matrix(Dis)
Dis<-Dis/1000
delta<- -mean(Dis)/log(0.01)
D<-exp(-Dis/delta)

#Create Similarity on use matrix from Balsa study#
n<-length(data$Balsa)
matrix<-matrix(NA,nrow=28,ncol=28)
for (i in 1:n){
for (k in i:n){
matrix[i,k]<-abs(data$Balsa[i]-data$Balsa[k])
}
}
SimB<-forceSymmetric(matrix)
SimB<-as.matrix(SimB)
zeta<- -mean(SimB)/log(0.01)
S<-exp(-SimB/zeta)
for (i in 1:n){
S[i,i]<-mean(S[i,-i])
}
```

```
#Create Adjacency matrix#
shp<-read.shp("portugalShapefile.shp")
dbf<-read.dbf("portugalShapefile.dbf")
data.combine<-combine.data.shapefile(data=data,shp=shp,dbf=dbf)
W.nb<-poly2nb(data.combine,row.names=rownames(data))
W.list<-nb2listw(W.nb,style="B")
W.mat<-nb2mat(W.nb,style="B")

#####################################################################
#Run Models using the prior (Best et.al. 1999)=>(0.001, 0.001)   #
#####################################################################
form_na<-data$cases_alah_na_00~data$c11manpc+ns(data$c11jovenspc,3)
form_na_full<-data$cases_alah_na_00~data$c11manpc
+ns(data$c11jovenspc,3)+ns(data$Balsapc,3)

#GAM in BYM with S matrix, calculated using the information from Balsa#
modelS<- S.CARbym(formula=form_na, family="poisson", data=data, W=S,
burnin=10000,n.sample=100000,thin=10,prior.tau2=c(0.1,0.1),
prior.sigma2=c(0.1,0.1))
modelS

#GAM in BYM with W matrix #
modelW<- S.CARbym(formula=form_na_full, family="poisson", data=data,
W=W.mat, burnin=10000,n.sample=100000,thin=10,
prior.tau2=c(0.1,0.1),prior.sigma2=c(0.1,0.1))
modelW

#GAM in BYM with Distance matrix #
modelD<- S.CARbym(formula=form_na_full, family="poisson", data=data,
W=D,  burnin=10000,n.sample=100000,thin=10,
prior.tau2=c(0.001,.001),prior.sigma2=c(0.001,0.001))
modelD
```

Appendix B

Data

This appendix presents the data used in this work.

	NUTS3 code	Description
1	111	Minho-Lima
2	112	Cávado
3	113	Ave
4	114	Grande Porto
5	115	Tâmega
6	116	Entre Douro e Vouga
7	117	Douro
8	118	Alto Trás-os-Montes
9	150	Algarve
10	161	Baixo Vouga
11	162	Baixo Mondego
12	163	Pinhal Litoral
13	164	Pinhal Interior Norte
14	165	Dão-Lafões
15	166	Pinhal Interior Sul
16	167	Serra da Estrela
17	168	Beira Interior Norte
18	169	Beira Interior Sul
19	16A	Cova da Beira
20	16B	Oeste
21	16C	Médio Tejo
22	171	Grande Lisboa
23	172	Península de Setúbal
24	181	Alentejo Litoral
25	182	Alto Alentejo
26	183	Alentejo Central
27	184	Baixo Alentejo
28	185	Lezíria do Tejo

Table B.1: From left to right: NUTS3 code, NUTS3 name.

		cases_AAD	AAD_expected	population
1	111	23469.96	17918	209885
2	112	12524.91	30083	326447
3	113	25191.37	38173	420726
4	114	117160.03	89699	1042208
5	115	3949.36	40647	437824
6	116	34817.76	20687	234278
7	117	7977.35	15141	175515
8	118	8220.02	15302	184285
9	150	20900.06	29629	351223
10	161	8730.01	28201	326804
11	162	29588.56	22967	277375
12	163	15745.44	18675	218423
13	164	0.00	9459	115102
14	165	20426.03	20700	241213
15	166	0.00	2771	35135
16	167	6733.02	3382	40816
17	168	5845.86	7630	93114
18	169	0.00	4951	62713
19	16A	2678.60	6358	76833
20	16B	50041.89	25373	297653
21	16C	34790.04	16180	193011
22	171	121306.70	139026	1646446
23	172	75997.59	54726	639697
24	181	0.00	6610	81169
25	182	9195.01	7988	99073
26	183	32100.12	11592	141660
27	184	11492.92	8714	105897
28	185	12655.12	17256	207281

Table B.2: From left to right: NUTS3 code, AAD observed number of cases, AAD expected number of cases, total population.

	man	young	cases_regularuse	Balsa
1	-1.99	-0.05	54176.73	0.71
2	-0.13	1.97	175217.49	0.56
3	0.38	1.42	177596.26	0.57
4	-0.91	0.85	858535.44	0.64
5	1.14	1.92	233757.22	0.63
6	0.65	0.89	113803.41	0.62
7	-0.22	-0.30	80826.19	0.60
8	0.48	-1.22	99430.98	0.60
9	1.33	0.52	186506.46	0.54
10	-0.07	0.64	122963.01	0.62
11	-1.23	0.03	168642.28	0.66
12	0.38	0.66	200007.30	0.64
13	-0.41	-0.90	21010.69	0.65
14	-0.75	-0.02	139820.07	0.58
15	-0.42	-2.09	10980.96	0.64
16	-0.95	-1.31	32285.13	0.70
17	-0.56	-1.13	54157.88	0.70
18	-0.38	-1.17	34283.69	0.64
19	-0.18	-0.73	2678.60	0.64
20	0.55	0.33	203200.88	0.64
21	-0.41	-0.43	121638.09	0.59
22	-1.53	1.09	1045814.08	0.64
23	-0.13	0.83	467994.93	0.62
24	2.95	-0.22	54905.31	0.60
25	0.33	-0.85	95581.57	0.22
26	0.20	-0.21	68468.28	0.59
27	1.60	-0.43	66384.76	0.56
28	0.29	-0.10	92731.15	0.60

Table B.3: From left to right: standardized covariates used: Proportion of men, proportion of people aged 18 to 34. Observed number of alcohol use cases (collected by WMHSI) and proportion of alcohol users as collected by Balsa et al. [4].

Appendix

Results

This appendix presents some of the results obtained in this work.

	SMRcrude_Low	SMRcrude	SMRcrude_High
1	0.72	1.29	2.34
2	0.16	0.43	1.16
3	0.30	0.66	1.47
4	0.71	1.28	2.31
5	0.02	0.11	0.76
6	1.03	1.70	2.83
7	0.24	0.58	1.40
8	0.18	0.48	1.28
9	0.32	0.71	1.59
10	0.11	0.35	1.08
11	0.73	1.33	2.39
12	0.39	0.82	1.73
13			
14	0.47	0.93	1.86
15			
16	1.18	1.93	3.15
17	0.33	0.73	1.63
18			
19	0.12	0.36	1.12
20	1.24	2.00	3.22
21	1.35	2.14	3.40
22	0.40	0.83	1.75
23	0.79	1.40	2.46
24			
25	0.58	1.11	2.14
26	1.86	2.80	4.22
27	0.74	1.34	2.42
28	0.32	0.72	1.61

Table C.1: Crude Standardized morbidity ratios and respective 95% CI.

	SMRSimilarity_Low	SMRSimilarity	SMRSimilarity_High
1	0.78	1.30	2.01
2	0.14	0.34	0.70
3	0.33	0.60	1.03
4	0.75	1.19	1.85
5	0.06	0.20	0.46
6	0.93	1.44	2.17
7	0.40	0.73	1.22
8	0.28	0.58	1.07
9	0.43	0.79	1.34
10	0.31	0.61	1.05
11	0.80	1.29	1.97
12	0.52	0.90	1.44
13	0.31	0.88	2.30
14	0.59	1.00	1.57
15	0.06	0.62	6.12
16	1.05	1.64	2.48
17	0.43	0.78	1.30
18	0.28	0.83	2.44
19	0.28	0.57	1.00
20	1.16	1.76	2.58
21	1.21	1.84	2.70
22	0.48	0.84	1.38
23	0.80	1.26	1.93
24	0.23	0.94	3.37
25	0.58	1.12	1.91
26	1.64	2.40	3.43
27	0.74	1.22	1.90
28	0.48	0.85	1.37

Table C.2: GRF similarity-based model posterior median SMRs and corresponding low and high SMRs for 90% of the posterior samples.

	SMRAdjacency_Low	SMRAdjacency	SMRAdjacency_High
1	0.79	1.32	2.07
2	0.14	0.35	0.72
3	0.31	0.60	1.04
4	0.70	1.15	1.81
5	0.05	0.17	0.44
6	0.88	1.40	2.15
7	0.34	0.66	1.12
8	0.25	0.55	1.04
9	0.46	0.85	1.45
10	0.27	0.56	0.98
11	0.76	1.25	1.95
12	0.50	0.88	1.44
13	0.27	0.86	2.39
14	0.56	0.99	1.59
15	0.07	0.97	16.61
16	1.06	1.70	2.57
17	0.43	0.80	1.35
18	0.22	0.81	2.75
19	0.24	0.52	0.96
20	1.15	1.77	2.63
21	1.22	1.87	2.75
22	0.49	0.88	1.47
23	0.82	1.31	2.05
24	0.17	0.82	3.83
25	0.57	1.06	1.78
26	1.70	2.48	3.52
27	0.76	1.25	1.96
28	0.49	0.87	1.41

Table C.3: GMRF adjacency-based model posterior median SMRs and corresponding low and high SMRs for 90% of the posterior samples.

	SMRDistance_Low	SMRDistance	SMRDistance_High
1	0.80	1.34	2.09
2	0.14	0.34	0.70
3	0.33	0.61	1.05
4	0.72	1.18	1.83
5	0.05	0.18	0.46
6	0.92	1.44	2.20
7	0.40	0.72	1.21
8	0.29	0.61	1.09
9	0.47	0.87	1.47
10	0.30	0.60	1.03
11	0.79	1.27	1.97
12	0.50	0.87	1.41
13	0.30	0.85	2.43
14	0.60	1.02	1.63
15	0.06	0.89	11.13
16	1.04	1.65	2.50
17	0.44	0.81	1.38
18	0.26	0.85	2.59
19	0.27	0.55	0.99
20	1.13	1.74	2.60
21	1.20	1.83	2.71
22	0.48	0.84	1.38
23	0.78	1.25	1.93
24	0.19	0.85	3.31
25	0.57	1.05	1.79
26	1.62	2.39	3.45
27	0.74	1.21	1.90
28	0.47	0.84	1.36

Table C.4: GMRF distance-based model posterior median SMRs and corresponding low and high SMRs for 90% of the posterior samples.